

**UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE CIÊNCIAS MATEMÁTICAS E DE COMPUTAÇÃO**

Maria Eugênia Ferreira da Fonseca

Precificação de imóveis Airbnb da cidade do Rio de Janeiro

**São Carlos
2020**

Maria Eugênia Ferreira da Fonseca

Precificação de imóveis Airbnb da cidade do Rio de Janeiro

Trabalho de conclusão de curso apresentado ao Centro de Ciências Matemáticas Aplicadas à Indústria do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, como parte dos requisitos para conclusão do MBA em Ciências de Dados.

Área de concentração: Ciências de Dados

Orientador: Prof. Dr. Gleici da Silva Castro Perdoná

Versão original

**São Carlos
2020**

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

F676p Fonseca, Maria Eugênia Ferreira da
Precificação de imóveis Airbnb da cidade do Rio
de Janeiro / Maria Eugênia Ferreira da Fonseca;
orientadora Gleici da Silva Castro Perdoná. -- São
Carlos, 2020.
60 p.

Trabalho de conclusão de curso (MBA em Ciência
de Dados) -- Instituto de Ciências Matemáticas e de
Computação, Universidade de São Paulo, 2020.

1. Precificação. 2. Airbnb. 3. Aluguel
temporário. I. Perdoná, Gleici da Silva Castro,
orient. II. Título.

Folha de aprovação em conformidade
com o padrão definido
pela Unidade.

No presente modelo consta como
folhadeaprovacao.pdf

AGRADECIMENTOS

Agradeço aos meus pais pelo amor, confiança e por sempre me apoiarem e acreditarem nos meus sonhos, tornando possível essa etapa. À minha namorada, que acompanhou de perto muitos finais de semana de estudo. Agradeço à Professora Gleici, que foi muito gentil e disponível para me orientar.

À todos que me apoiaram e contribuíram direta ou indiretamente para minha formação, meu muito obrigada!

*“If you can meet with Triumph and Disaster
And treat those two impostors just the same.”*

Rudyard Kipling

RESUMO

Fonseca, M. E. F. **Precificação de imóveis Airbnb da cidade do Rio de Janeiro.** 2020. 60p. Monografia (MBA em Ciências de Dados) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2020.

O Airbnb é um dos tipos de serviços da chamada economia compartilhada, que é uma proposta de consumo mais sustentável através do compartilhamento de bens e serviços. Através do site pessoas podem obter um dinheiro extra ao alugar recursos que estão subutilizados, como um quarto parado em casa. O objetivo deste trabalho é de entender a precificação dos anúncios de acomodações do Airbnb na cidade do Rio de Janeiro e de identificar padrões ou fatores que possam levar um imóvel a ter um aluguel mais elevado. Utilizou-se regras de associação para entender o perfil dos imóveis, visualizações espaciais para entender a localidade dos imóveis, processamento de texto para comentários e aplicou-se técnicas de aprendizado de máquina para a tarefa de precificação. Nos resultados da precificação, algumas dificuldades surgiram devido à complexidade e heterogeneidade dos dados, mas ainda assim obteve-se resultados comparáveis com a literatura disponível.

Palavras-chave: Precificação. Airbnb. Aluguel temporário. Aprendizado de máquina. Regras de associação.

ABSTRACT

Fonseca, M. E. F. **Pricing of Airbnb properties in the city of Rio de Janeiro.** 2020. 60p. Monografia (MBA em Ciências de Dados) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2020.

Airbnb is one type of service in the so-called sharing economy, which is a more sustainable consumption proposal through the sharing of goods and services. Through the website, people can get extra money when renting underutilized resources, such as a room at home. The objective of this work is to understand the pricing of Airbnb accommodation ads in the city of Rio de Janeiro and to identify patterns or factors that could lead a property to have a higher rent. Association rules were used to understand the profile of the properties, spatial views to understand the location of the properties, word processing for comments and machine learning techniques were applied for the pricing task. In the pricing results, some difficulties arose due to the complexity and heterogeneity of the data, but results were still comparable to the available literature.

Keywords: Pricing. Airbnb. Short-term rental. Machine learning. Association rules.

LISTA DE FIGURAS

Figura 1 – Comparação entre floresta aleatória e <i>gradient boosting</i>	27
Figura 2 – Histograma do preço	31
Figura 3 – Boxplot do preço por tipo de acomodação	31
Figura 4 – Boxplot do preço pela quantidade de hóspedes	32
Figura 5 – Histograma da disponibilidade de reserva	33
Figura 6 – Comodidades mais frequentes	34
Figura 7 – Comodidades menos frequentes	34
Figura 8 – Efeito positivo de comodidades na mediana do preço	35
Figura 9 – Efeito negativo de comodidades na mediana do preço	36
Figura 10 – Heatmap dos imóveis Airbnb na cidade do Rio de Janeiro	38
Figura 11 – Heatmap dos imóveis Airbnb para aluguel na cidade do Rio de Janeiro por preço	38
Figura 12 – Heatmap dos imóveis Airbnb para aluguel por preço na região do Leblon, Ipanema e Copacabana	39
Figura 13 – Heatmap dos imóveis Airbnb para aluguel por preço na região do Centro	39
Figura 14 – Principais pontos turísticos na cidade do Rio de Janeiro	40
Figura 15 – Nuvem de palavra dos nomes de anúncio em português dos imóveis . .	41
Figura 16 – Nuvem de palavra dos nomes de anúncio em inglês dos imóveis	41
Figura 17 – Nuvem de palavra para comentários em Português	42
Figura 18 – Nuvem de palavra para comentários de avaliações ruins em Português .	42
Figura 19 – Nuvem de palavra para comentários em Inglês	43
Figura 20 – As vinte variáveis mais importantes	48

LISTA DE TABELAS

Tabela 1 – Preço por tipo de acomodação	31
Tabela 2 – Preço por quantidade de hóspedes	32
Tabela 3 – Disponibilidade de reserva por faixas de preço	33
Tabela 4 – Resultados - Variáveis originais	43
Tabela 5 – Resultados - Variáveis com transformações	44
Tabela 6 – Resultados - Modelos com ajuste de hiperparâmetros	44
Tabela 7 – Resultados - Modelos após seleção de variáveis	45
Tabela 8 – Resultados - Modelos diferentes para cada tipo de acomodação	45
Tabela 9 – Tabela com todas as variáveis utilizadas	57

SUMÁRIO

1	INTRODUÇÃO	19
1.1	Introdução	19
1.2	Objetivos	22
1.3	Organização	22
2	METODOLOGIA	23
2.1	Pré-processamento	23
2.1.1	Separação da base de dados em treino e teste	23
2.1.2	Seleção de variáveis	24
2.1.3	Imputação	24
2.1.4	Tratamento de variáveis categóricas	24
2.1.5	Transformações em variáveis numéricas	25
2.1.6	Criação de novos atributos	25
2.2	Regras de associação	25
2.3	Modelos de regressão	26
2.3.1	Régressão linear	26
2.3.2	Régressão <i>Ridge</i>	26
2.3.3	Floresta aleatória	27
2.3.4	<i>Gradient boosting</i>	27
2.3.5	Validação cruzada e ajuste de hiperparâmetros	28
2.3.6	Critérios para seleção dos modelos	28
3	RESULTADOS	29
3.1	Descrição do pré-processamento	29
3.2	Análise descritiva e regras de associação	30
3.3	Localização dos imóveis	37
3.4	Análises dos atributos em formato texto	40
3.5	Resultados da precificação	43
4	DISCUSSÕES E CONCLUSÕES	47
4.1	Considerações finais e trabalhos futuros	49
	REFERÊNCIAS	53

ANEXOS **55**

ANEXO A – LISTA DE VARIÁVEIS UTILIZADAS 57

1 INTRODUÇÃO

1.1 Introdução

Economia compartilhada é uma nova forma de negócio, que traz um consumo mais sustentável através do compartilhamento de bens e serviços. Através de sites ou aplicativos, pessoas podem alugar de carros a apartamentos, utensílios de cozinha e até mesmo aviões executivos (BRASIL, 2016). Uma grande vantagem do ponto de vista de quem anuncia o compartilhamento de bens e serviços é poder utilizar recursos que estão subutilizados, como um quarto parado em casa. Do ponto de quem utiliza estes serviços, uma vantagem está em não ter que investir em adquirir bens para poder usufruir. Um exemplo de economia compartilhada é o Airbnb, um serviço online que permite as pessoas anunciar, descobrirem e reservarem acomodações e experiências em mais de 191 países. A plataforma é usada principalmente por turistas e profissionais em trabalho, tendo como vantagens os custos mais baixos e a facilidade de alugar um imóvel de forma totalmente digital e sem muita burocracia (GARRETT, 2017; AIRBNB, 2020a).

Segundo um estudo da PricewaterhouseCoopers) (2015) de 2015, era esperado que em 2025 a economia compartilhada gerasse 335 bilhões de dólares em receitas, principalmente nos setores de hospedagem, compartilhamento de veículos, finanças e plataformas de *streaming* de música e vídeo. Em um artigo publicado na Forbes em 2018 (TEAM, 2018), estimava-se que o Airbnb estava avaliado em 38 bilhões de dólares; agora, após surto do coronavírus, estima-se que a empresa esteja avaliada em 18 bilhões de dólares (EAGLESHAM, 2020). Para entender a magnitude da atuação do Airbnb, o mesmo estudo de 2015 (PRICEWATERHOUSECOOPERS), 2015) estimou que o serviço hospedava em média 425 mil hóspedes por noite, quase 22% a mais que a rede Hilton no mundo inteiro.

Existem dois ramos no Airbnb, o de acomodações e o de experiências. As experiências do Airbnb são atividades e passeios fora do usual, como aulas de dança e culinária com anfitriões de todo o mundo, de forma online ou presencial. Pelo ramo de acomodações, que será o foco deste trabalho, é possível fazer reservas em variados tipos de acomodação, desde quartos compartilhados a casas inteiras, e de imóveis simples a exóticos e luxuosos.

Para começar a usar o Airbnb, o anfitrião (pessoa que está anunciando o espaço) precisa inserir informações como a localização do seu espaço, o tipo de propriedade, o número de quartos e banheiros disponíveis, adicionar fotos do ambiente e escrever uma descrição do anúncio (AIRBNB, 2020b). O próximo passo é organizar a logística do anúncio, como adicionar regras da casa (restrições a fumo, animais ou festas, por exemplo) e configurar o calendário, para marcar as datas com disponibilidade para reserva. Finalmente, é preciso escolher o preço por noite. O próprio Airbnb tem ferramentas para

ajudar nesta tarefa, como o Preço Inteligente, que ajuda a definir os preços de acordo com a demanda, além de fornecer controles para preços personalizados para ocasiões como fins de semana (AIRBNB, 2020b). Ainda assim, segundo [Kalehbasti, Nikolenko e Rezaei \(2019\)](#), precificar o aluguel de um imóvel é uma tarefa desafiadora para os proprietários, pois afeta diretamente a quantidade de hóspedes que irão escolher o local.

A economia compartilhada tem vivenciado um crescimento explosivo em anos recentes. Especialistas ([BRASIL, 2016](#)) apontam para dois fatores que explicam este crescimento. O primeiro, a crise econômica de 2008, quando muitos países enfrentaram recessões. O segundo motivo, os avanços tecnológicos e investimentos massivos: a tecnologia agora proporcionou o contato facilitado entre quem precisa de um bem ou serviço de quem pode oferece-lo. Como o tema é recente, a área de estudo é relativamente nova e, especificamente no caso de alugueis temporários, a literatura encontrada aponta que existe espaço para inclusão de análises que levem a melhor entender o comportamento dos usuários, para por exemplo, precificar os alugueis. Na literatura pesquisada foram encontradas referências de precificação de Airbnb em cidades como Nova Iorque ([KALEHBASTI; NIKOLENKO; REZAEI, 2019; LUO; ZHOU; ZHOU, 2019](#)), Paris e Berlim ([LUO; ZHOU; ZHOU, 2019](#)), algumas cidades da Alemanha ([LUO; ZHOU; ZHOU, 2019](#)), entre outras, mas não foi encontrada nenhuma publicação com precificação no Rio de Janeiro ou em qualquer outra cidade brasileira.

[Kalehbasti, Nikolenko e Rezaei \(2019\)](#) usam métodos como regressão linear, modelos baseados em árvores, SVR (*Support-Vector Regression*), clusterização usando k-médias e redes neurais em uma base de dados sobre Airbnb da cidade de Nova Iorque. Os autores também fazem uma análise de sentimento em uma informação textual (avaliações dos hóspedes) para complementar o modelo de precificação, além de usar técnicas como a regressão Lasso para a redução de variáveis. Os autores encontraram que o melhor modelo foi o SVR, com um R^2 de 0,7768 na base de treino e 0,6901 na base de teste.

Em 2019, [Luo, Zhou e Zhou \(2019\)](#) estudaram como um modelo de precificação de Airbnb poderia ser generalizado para mais cidades. As cidades consideradas no estudo foram Nova Iorque e Paris, com o objetivo de prever a precificação de Berlim. Os autores mostraram que o modelo treinado no conjunto de dados combinado de Nova Iorque e Paris, ao invés dos conjuntos de dados individuais, é mais generalizável para prever o preço em uma nova cidade. O melhor modelo, rede neural, obteve um R^2 de 0,816 na base de treino e 0,773 na base de teste. Esse estudo dá indícios de que modelos e técnicas utilizadas em outras cidades possam servir de guia para o estudo no contexto da cidade do Rio de Janeiro.

No estudo de [Teubner, Hawlitschek e Dann \(2017\)](#) foram avaliados fatores que contribuem para a precificação do aluguel temporário em 86 cidades alemãs, principalmente a reputação do anfitrião. Cinco categorias de atributos foram consideradas: reputação,

atributos do apartamento, atributos da cidade, conveniência e características sobre o proprietário. Esse estudo trouxe duas novidades em relação à maioria dos trabalhos sobre o tema. Primeiro, acrescentou uma informação sobre a distância do imóvel até o centro da cidade utilizando coordenadas geográficas. E segundo, fez uma análise mais aprofundada sobre o anfitrião: considerou o gênero, classificado manualmente a partir do nome, e informações a partir da foto, usando uma API de inteligência artificial da Microsoft para estimativa da idade e grau do sorriso.

Além dos estudos diretamente relacionados à precificação, encontramos artigos que avaliam questões mais sociais, como preconceito dentro da plataforma ([EDELMAN; LUCA, 2014](#); [EDELMAN; LUCA; SVIRSKY, 2017](#)) e os reais benefícios desse tipo de empreendimento para a comunidade local ([BARRON; KUNG; PROSERPIO, 2018](#)).

"Em um esforço para facilitar a confiança, muitas plataformas incentivam anfitriões a fornecerem perfis pessoais e até postar fotos deles mesmos. Contudo, esses recursos também podem facilitar a discriminação com base na raça, sexo, idade ou outros aspectos da aparência"([EDELMAN; LUCA, 2014](#)). Esse estudo mostrou que proprietários não negros conseguem cobrar aproximadamente 12% a mais, considerando características similares de imóveis. Um outro estudo dos mesmos autores ([EDELMAN; LUCA; SVIRSKY, 2017](#)) sugere que "solicitações de hóspedes com nomes distintamente afro-americanos são aproximadamente 16% menos prováveis de serem aceitos do que solicitações de hóspedes idênticos com nomes distintamente brancos".

Além de estudos que avaliam a questão do preconceito no Airbnb, existem também questionamento quanto aos reais benefícios da plataforma para as cidades. Um estudo de 2018 ([BARRON; KUNG; PROSERPIO, 2018](#)) avalia os efeitos do Airbnb em preço de imóveis e alugueis não temporários. As críticas apontam que apesar da popularidade entre os consumidores esse tipo de comércio é mal regulamentado. O estudo indica que existe uma relação entre a oferta de Airbnb em um determinado local e um aumento no mercado imobiliário ao que se refere o valor das propriedades e dos alugueis não temporários.

Todos estas questões originaram na disponibilização da base de dados pelo site *Inside Airbnb*, com o objetivo de fornecer publicamente os dados do Airbnb para que as pessoas possam entender como esse mercado está afetando a comunidade local, entre outras problemas.

Este trabalho tem como objetivo propor modelos de predição para precificação de aluguéis temporários na cidade de Rio de Janeiro. Desta forma, contribuir para o conhecimento de precificação de um imóvel no local e compreensão do comportamento do anfitrião na precificação do mesmo.

1.2 Objetivos

Este trabalho propõe-se a desenvolver precificação aplicando técnicas de aprendizado de máquina em dados de reserva de acomodações do Airbnb na cidade do Rio de Janeiro.

Os objetivos específicos são:

1. Estudar os modelos de predição mais comumente empregados em precificação de aluguéis temporários;
2. Comparar a performance de modelos de regressão em um conjunto de dados real do Airbnb, utilizando o Python, com o uso de medidas de avaliação da capacidade preditiva;
3. Identificar padrões ou fatores que possam levar um imóvel a ter uma aluguel mais caro.
4. Avaliar possíveis impactos de atributos externos, como questões de gênero que foram levantados nas referências bibliográficas.

1.3 Organização

No Capítulo 2, apresenta-se a metodologia utilizada, bem como os conceitos e explicação de cada algoritmo e técnica utilizada. No Capítulo 3 é apresentada uma explicação das variáveis disponíveis, juntamente com análises descritivas e os resultados da modelagem. No Capítulo 4 são discutidas as comparações entre os modelos considerados e as conclusões.

2 METODOLOGIA

Este trabalho propõe-se a estudar as bases de dados disponibilizadas pelo site *Inside Airbnb* para realizar uma tarefa de precificação de aluguéis temporários do Airbnb. Atualmente estão disponibilizadas informações de locação de mais de 90 cidades, dentre as quais quatro cidades da América Latina e apenas uma no Brasil, na cidade do Rio de Janeiro.

No site estão disponíveis dados mensais da cidade do Rio de Janeiro de Abril de 2018 a Outubro de 2020, com algumas falhas na periodicidade mensal de atualização. Cada uma destas bases contêm três tabelas:

- Tabela *listings*, que apresenta as principais informações sobre o Airbnb a ser alugado. Como informações sobre o imóvel, preço e anfitrião.
- Tabela *calendar*, que apresenta o preço e a disponibilidade de cada imóvel para os próximos 365 dias.
- Tabela *reviews*, que contém as avaliações (em texto) de todos os imóveis e em diversos idiomas.

O foco do projeto será usar a primeira tabela, *listings*, para ajustar um modelo de predição do preço dos alugueis. Escolhemos a base de dados do mês de Janeiro de 2020, antes de um possível efeito no preço causado pelo carnaval, que em 2020 aconteceu no final do mês de fevereiro e atraiu mais de 2,1 milhões turistas para a cidade e mais de 10 milhões de pessoas circulando, segundo a prefeitura do município do Rio de Janeiro ([JANEIRO, 2020](#)); e pelo Coronavírus, que teve o primeiro caso confirmado no Brasil também no final do mês de fevereiro ([PAULO, 2020](#)). A base de dados considerada (tabela *listings*) traz informações de 34.754 imóveis e contém 105 variáveis. Uma Analise descritiva inicial apresentará mais detalhes sobre a fonte de dados.

2.1 Pré-processamento

Antes da etapa de modelagem e utilização de algorítimos de *machine learning* para precificação, é necessário passar por processos de pré-processamento de dados, que serão descritos nesta seção e que, no caso deste trabalho, envolvem etapas como o tratamento de valores faltantes e seleção de variáveis.

2.1.1 Separação da base de dados em treino e teste

"Aprender os parâmetros de uma função de previsão e testá-los com os mesmos dados é um erro metodológico: um modelo que apenas repetisse os rótulos das amostras

que acabou de ver teria uma pontuação perfeita, mas não conseguiria prever nada de útil em dados ainda não vistos. Essa situação é chamada de *overfitting*. Para evitá-lo, é prática comum, ao realizar um experimento de aprendizado de máquina (supervisionado), separar parte dos dados disponíveis como um conjunto de teste" (PEDREGOSA et al., 2011; LEARN, 2020b). Dessa forma, iremos separar o nosso banco de dados em treino e teste.

2.1.2 Seleção de variáveis

Quando construindo um modelo de regressão a seleção de variáveis é um problema que aparece com frequência. Alguns métodos descritos na literatura são os baseados em testes estatísticos, como o *stepwise regression*, e outros da categoria de regularização, como o modelo Ridge (DESBOULETS, 2018).

Utilizaremos o modelo Ridge, que será explicado em maiores detalhes em uma seção posterior. Além disso, outras variáveis foram removidas pelos seguintes critérios: valores constantes, informações de URL, algumas datas, variáveis textuais ou com mais de 90% de dados faltantes.

2.1.3 Imputação

Dados faltantes podem ser entendidos como valores ausentes para uma determinada variável. O problema da falta de dados é relativamente comum em bases de dados do mundo real e são um problema para a modelagem (PEDREGOSA et al., 2011; LEARN, 2020e). Para este trabalho, iremos usar a biblioteca *scikit-learn* da linguagem de programação Python, que assume que todos os valores fornecidos são numéricos e completos. Uma estratégia básica para usar conjunto de dados incompletos é descartar observações que contém valores faltantes, mas isso acarreta na perda de dados que podem ser importantes, ainda mais se implicar no descarte de muitas observações. Uma estratégia melhor é imputar os valores ausentes, por exemplo, inferindo os valores a partir das informações presentes. Um método básico de imputação é considerar a média ou mediana para variáveis numéricas; sendo a segunda opção mais apropriada para variáveis de distribuição assimétrica.

2.1.4 Tratamento de variáveis categóricas

Como mencionado na sessão anterior, a bilbioteca *scikit-learn* assume que todos os valores fornecidos são numéricos. Dessa forma, implica-se a fazer um tratamento quanto às variáveis categóricas. Optamos por transformar as variáveis categóricas nominais em variáveis *dummy*, que são uma representação com atributos binários, e transformar as variáveis categóricas ordinais em numéricas, de acordo com a ordem crescente original.

2.1.5 Transformações em variáveis numéricas

Variáveis com distribuições assimétricas, de cauda longa ou com *outliers* podem ser um problema para a modelagem. O uso de transformações é uma técnica comum neste caso. Testaremos o uso do logaritmo natural para transformações em variáveis assimétricas, incluindo a variável resposta preço.

2.1.6 Criação de novos atributos

Como no estudo de [Teubner, Hawlitschek e Dann \(2017\)](#), também vamos construir um atributo com a classificação do sexo do anfitrião a partir do seu nome. Usamos o pacote *genderBR*, disponível na linguagem R, que usa dados do Censo de 2010 do IBGE para inferir o sexo de uma pessoa. Segundo o autor do pacote, [Meireles \(2017\)](#), "a precisão média do método sempre foi maior que 95%, sem sinal de viés".

Além do atributo sexo, vamos gerar informações a partir da tabela *reviews*. Uma hipótese que queremos testar é se imóveis com mais avaliações em outros idiomas são usualmente especificados com maior valor. Para isso, fizemos o reconhecimento de idioma a partir das avaliações usando a biblioteca *langdetect*, disponível em Python.

Por fim, queremos entender se a localização dos imóveis relativa aos pontos turísticos da cidade do Rio de Janeiro é algo importante para a especificação. Construímos uma lista a partir de atrações classificadas pelo TripAdvisor, que inclui avaliações dos viajantes, pontuações, fotos e popularidade. A partir desta lista, obtemos a latitude e longitude de cada ponto e calculamos a distância para a latitude e longitude do imóvel, utilizando a biblioteca *geodesic*, disponível em Python, que considera a curvatura da terra para o cálculo de distância.

2.2 Regras de associação

[Borgelt e Kruse \(2002\)](#) definem regras de associação como uma técnica poderosa para a chamada análise de carrinho de supermercado, que visa encontrar regularidades no comportamento de compra dos clientes de supermercados, lojas on-line e similares. Com as regras de associação tenta-se encontrar conjuntos de itens que são frequentemente comprados juntos, de forma que a partir da presença de determinados produtos em um carrinho de compras se possa inferir, com uma probabilidade, que determinados outros produtos estão presentes. Esse tipo de informação pode ajudar na tomada de decisão para, por exemplo, organizar promoções ou a disposição dos itens na loja.

Uma das dificuldades com as regras de associação são o grande número de regras que podem ser geradas, onde para um certo número de itens podem existir muitas combinações diferentes. [Borgelt e Kruse \(2002\)](#) explicam que a importância de uma regra normalmente é mensurada de duas formas: pelo suporte e pela confiança. O suporte é o percentual de

transações onde tal regra ocorreu. Já a confiança, é o número de casos onde a regra está correta relativa ao número de casos que ela é aplicável. Por exemplo, supondo uma regra de que quem compra pão e manteiga também compra leite ([AGRAWAL; IMIELIŃSKI; SWAMI, 1993](#)) tem um suporte de 5% e uma confiança de 90%, temos as seguintes interpretações: em 5% das compras do supermercado temos essa regra acontecendo e que se pão e manteiga foram comprados, temos 90% de chance de ter leite junto.

As regras de associação serão usadas no contexto de comodidades, que são atributos relativos ao imóvel, como a presença de cozinha, ar-condicionado e estacionamento. Utilizamos a implementação do algoritmo *apriori*, descrito no artigo de [Borgelt e Kruse \(2002\)](#).

2.3 Modelos de regressão

Um modelo de regressão avaliar a relação entre uma variável resposta e variáveis explicativas. Selecionada a variável resposta preço, diferentes modelos de regressão foram considerados. A modelagem foi realizada usando o Scikit-learn, disponível em Python ([PEDREGOSA et al., 2011](#)), e detalhes metodológicos dos modelos foram obtidos na documentação oficial da biblioteca.

2.3.1 Regressão linear

Regressão linear pode ser descrita como um método de regressão onde a variável resposta pode ser modelada por uma combinação linear das variáveis explicativas ([PEDREGOSA et al., 2011; LEARN, 2020a](#)). Matematicamente podemos escrever:

$$\hat{y}(w, x) = w_0 + w_1x_1 + \dots + w_px_p$$

Onde $w = (w_1, \dots, w_p)$ são os coeficientes do modelo e w_0 é o intercepto. Para ajustar o modelo de regressão linear procuramos os coeficientes que minimizem a soma residual dos quadrados entre os dados e os valores previstos pela aproximação linear. Algumas suposições do modelo são de variáveis explicativas independentes e de que os erros são observações independentes e seguem uma distribuição normal com média zero e variância σ^2 . Matematicamente, o modelo tenta resolver o seguinte mínimo:

$$\min_w \|Xw - y\|_2^2$$

2.3.2 Regressão *Ridge*

A regressão *Ridge* é uma extensão da regressão linear, com imposição de penalidade para o tamanho dos coeficientes ([PEDREGOSA et al., 2011; LEARN, 2020a](#)). Os coeficientes Rige minimizam:

$$\min_w \|Xw - y\|_2^2 + \alpha \|w\|_2^2$$

Onde o parâmetro α é um coeficiente de complexidade que controla a penalidade do modelo: quanto maior o valor de α , maior é a penalidade.

2.3.3 Floresta aleatória

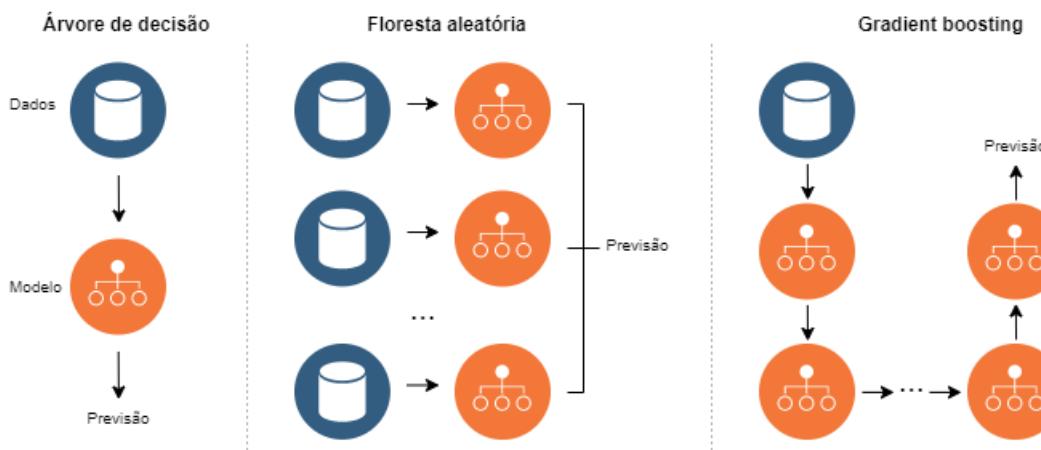
Floresta aleatória é um modelo que cria uma combinação, ou *ensemble*, de várias árvores de decisão em várias subamostras do conjunto de dados. O modelo então usa a média dessas árvores para tentar melhorar a previsão e para tentar controlar o *overfitting* (PEDREGOSA et al., 2011; LEARN, 2020c). A aleatoriedade do modelo está no uso de subamostras para cada árvore e no uso de subconjuntos aleatórios das variáveis explicativas para a construção dos nós.

2.3.4 Gradient boosting

O modelo *gradient boosting* é uma técnica de aprendizado supervisionado, que produz um modelo na forma de uma combinação de modelos de previsão mais fracos (PEDREGOSA et al., 2011; LEARN, 2020d). Segundo (FRIEDMAN, 2001), o modelo *gradient boosting* em árvores de decisão produz procedimentos competitivos, altamente robustos e interpretáveis, tanto para regressão e classificação, e especialmente apropriados para minerar dados menos limpos.

Como ilustrado na Figura 1, enquanto a floresta aleatória ajusta árvores de decisão em paralelo e pondera os resultados, o *gradient boosting* ajuste as árvores em sequência, modelando especialmente os resíduos das árvores anteriores.

Figura 1 – Comparaçāo entre floresta aleatória e *gradient boosting*



2.3.5 Validação cruzada e ajuste de hiperparâmetros

Os hiperparâmetros dos modelos não fazem parte do aprendizado do algoritmo, sendo definido previamente pelo usuário. Para avaliar modelos com diferentes hiperparâmetros não podemos usar a base de teste, pois queremos manter estes dados apenas para a validação final. Por isso, é utilizada a validação cruzada para o processo de ajuste dos hiperparâmetros. Uma abordagem é usar a validação K-fold, que separa o conjunto de treino em k pedaços. Para cada um dos k pedaços treinamos o modelo nos $k-1$ pedaços restantes e testamos o modelo no pedaço k selecionado.

2.3.6 Critérios para seleção dos modelos

Iremos fazer o ajuste de algoritmos diferentes e para poder compará-los, iremos usar duas métricas como critério: a raiz do erro quadrado médio (RMSE, *root-mean-square error* em inglês) e o coeficiente de determinação R^2 . O RMSE é uma medida de distância entre os valores estimados pelo modelo e os valores observados. Quanto maior o valor, pior. Já o R^2 , varia entre 0 e 1, e pode ser entendido como a quantidade da variância dos dados que é explicada pelo modelo. Sendo assim, quanto maior o seu valor, melhor a explicação da variância dos dados pelo modelo.

3 RESULTADOS

Antes da etapa de modelagem e utilização de algoritmos de *machine learning* precisamos passar por processos de pré-processamento de dados, para o ajuste dos dados no formato necessário, e por uma análise descritiva, para conhecermos melhor os dados e formular hipóteses. Ambos assuntos serão tratados neste capítulo.

A principal base de dados considerada (tabela *listings*) traz informações de 34.754 imóveis e contém 105 variáveis. Ao total, as informações foram classificadas em cinco temas e trazem informação como as seguintes:

- Anfitrião: tempo de resposta e a quantidade de outros anúncios pelo mesmo anfitrião no Airbnb.
- Disponibilidade, requisitos e preços: preço por diária (nossa variável resposta), número mínimo de diárias, política de cancelamento e dias disponíveis para aluguel.
- Imóvel: tipo de acomodação (se é espaço inteiro ou compartilhado), número de quartos e tipos de comodidade (como ar-condicionado, Wi-Fi e piscina).
- Localização: nome do bairro e distâncias em quilômetros do imóvel até os vinte principais pontos turísticos do Rio de Janeiro.
- Reputação: notas e quantidade de comentários do imóvel.

Para mais detalhes sobre todas as variáveis utilizadas, segue tabela completa no Anexo [4.1](#).

3.1 Descrição do pré-processamento

Após análises iniciais, identificamos variáveis com muitos valores faltantes, valores constantes (como o nome do país), cardinalidade alta (como nome de bairro), distribuição assimétrica ou com *outliers* (como o número mínimo de diárias, que está concentrado em 1 ou 2 dias, mas tem um valor máximo de 1122 dias).

Separamos a nossa base de dados em dois grupo, como motivado pela subseção [2.1.1](#). O primeiro grupo sendo a base de treino e composto por 70% dos dados originais, responsável pelo aprendizado do modelo. Já o segundo grupo, base de teste, composto pelos restantes 30% e utilizado para avaliar o desempenho do modelo com dados novos.

Dentre as variáveis consideradas viáveis para a análise de precificação, observamos muitos valores faltantes; atributos continham entre 0,16% e 44% de informações nulas.

Como a maioria das variáveis na nossa base de dados segue uma distribuição assimétrica, optamos por imputar a mediana. Importante destacar que a mediana considerada para imputação tanto na base de treino como teste foi obtida apenas a partir da base de dados de treino. A exceção foi a variável de número de avaliações mensais, que possuía 41,5% de dados faltantes. Para esta variável, entendemos que o valor faltante significa zero avaliações.

Como explicado pela subseção 2.1.4, precisávamos tratar as variáveis categóricas. A grande maioria das variáveis em questão eram categóricas nominais, sendo transformadas em variáveis *dummy*. A única exceção sendo uma variável de tempo de resposta do anfitrião, onde existia ordem entre as categorias; para este caso transformamos as categorias em números de acordo com a ordem crescente original.

Ao fazer o tratamento das variáveis categóricas passamos a ter muitos atributos, principalmente por causa de variáveis com alta cardinalidade. Para informação do bairro, por exemplo, existem mais de 150 valores distintos e para tipos de comodidade, como ar-condicionado e cozinha, mais de 170. Limitamos os níveis de categorias para apenas aqueles que são frequentes em pelo menos 5% dos imóveis.

Por fim, criamos vinte novos atributos, que são a distância em quilômetros de cada imóvel às vinte atrações turísticas consideradas. Selecionadas as variáveis factíveis, realizadas as devidas imputações, inserindo os atributos novos e feito o tratamento das variáveis categóricas, obtemos uma base de dados com 34.754 observações e 188 colunas.

3.2 Análise descritiva e regras de associação

A seguir serão apresentadas análises descritivas das variáveis que julgamos as mais importantes para explicar o problema de precificação.

Ao analisar o atributo preço, nossa variável resposta, percebemos como os valores são assimétricos à direita (Figura 2). Enquanto pouco mais da metade dos imóveis tem uma diária no valor de 300 reais, existem valores de diárias de quase 42 mil reais. Para facilitar a visualização, apresentaremos gráficos de preço limitados a 4 mil reais, o que já representa mais de 98% dos imóveis.

Os espaços alugados podem ser de quatro tipos: espaços inteiros (71,6%), quartos inteiros (25,6%), quartos compartilhados (2,1%) e quartos de hotel (0,7%) (Tabela 1). Acreditamos que essa variável deve afetar o preço e, como pode ser visto na Tabela 1 e na Figura 3), o tipo espaço inteiro realmente parece ter um preço superior aos outros. Para todos os tipos de acomodação notam-se muitos outliers.

Figura 2 – Histograma do preço

Histograma do Preço
Visão limitada a preços de até R\$ 4.000 (98% dos imóveis)

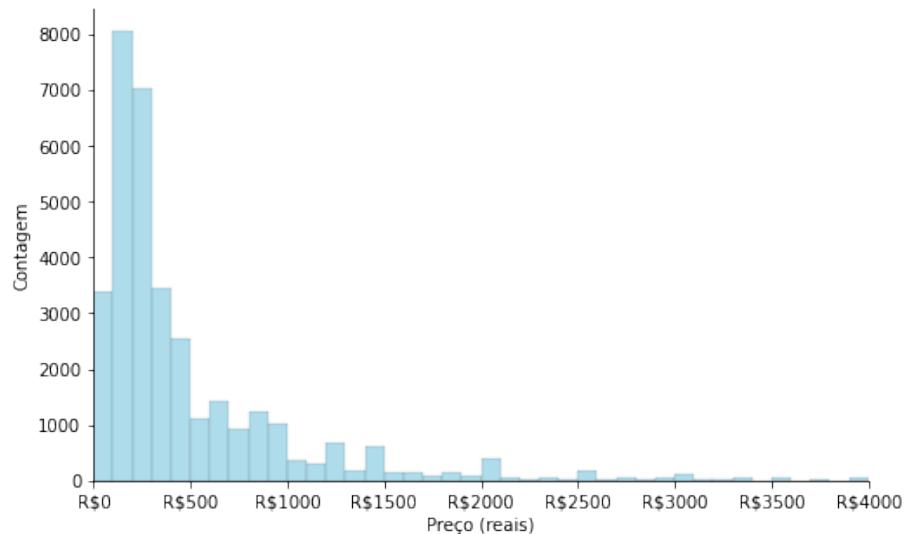
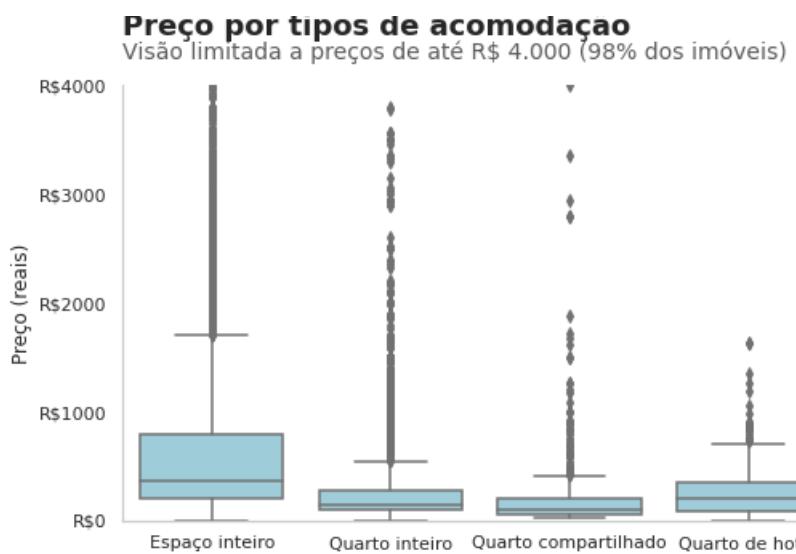


Tabela 1 – Preço por tipo de acomodação

Tipo de acomodação	Preço (reais)				Total
	0-200	201-500	501-1000	Mais de 1000	
Espaço inteiro	22%	40%	20%	17%	72%
Quarto inteiro	61%	29%	7%	3%	26%
Quarto compartilhado	73%	19%	5%	3%	2%
Quarto de hotel	51%	31%	10%	9%	1%
Total	34%	37%	16%	13%	100%

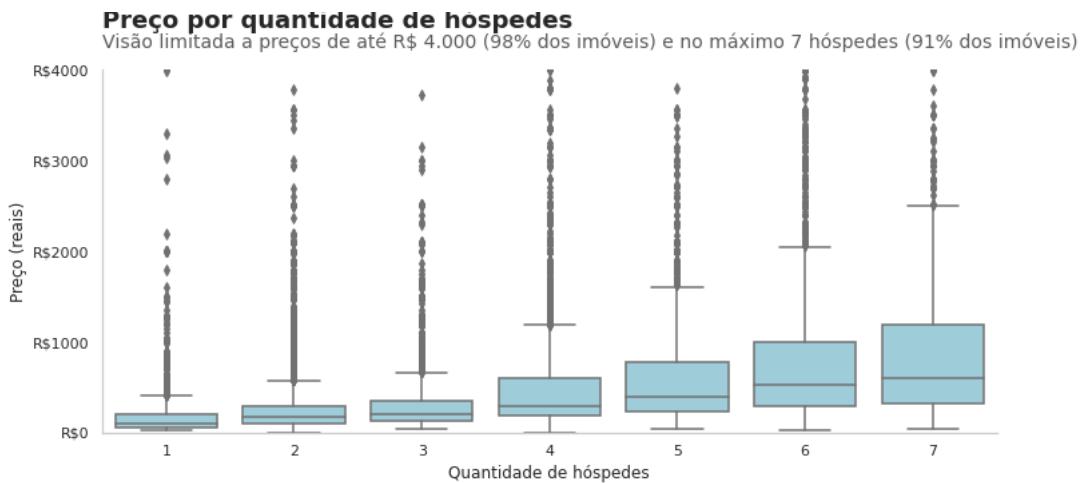
Figura 3 – Boxplot do preço por tipo de acomodação



Cada espaço pode receber uma certa quantidade de hóspedes, sendo o mínimo

uma pessoa. A variável segue uma distribuição assimétrica à direita: cerca de 30% dos imóveis acomoda até 2 pessoas, quase 67% acomoda até 4 pessoas e existe um anúncio com hospedagem para 160 pessoas. A hipótese a ser estudada é que a quantidade de hóspedes que um espaço pode receber é diretamente relacionada com o preço. Pela Figura 4 percebemos que o preço parece sim seguir a tendência de aumento da quantidade de hóspedes.

Figura 4 – Boxplot do preço pela quantidade de hóspedes



Como pode ser visto pela Tabela 2, preço e números de hóspedes realmente andam junto. Dos espaços de até duzentos reais a diária, 53% acomodam até duas pessoas; e dos espaços que custam mais de mil reais, 40% ocupam de 5 a 7 hóspedes e 31% aceitam mais de 7 pessoas.

Tabela 2 – Preço por quantidade de hóspedes

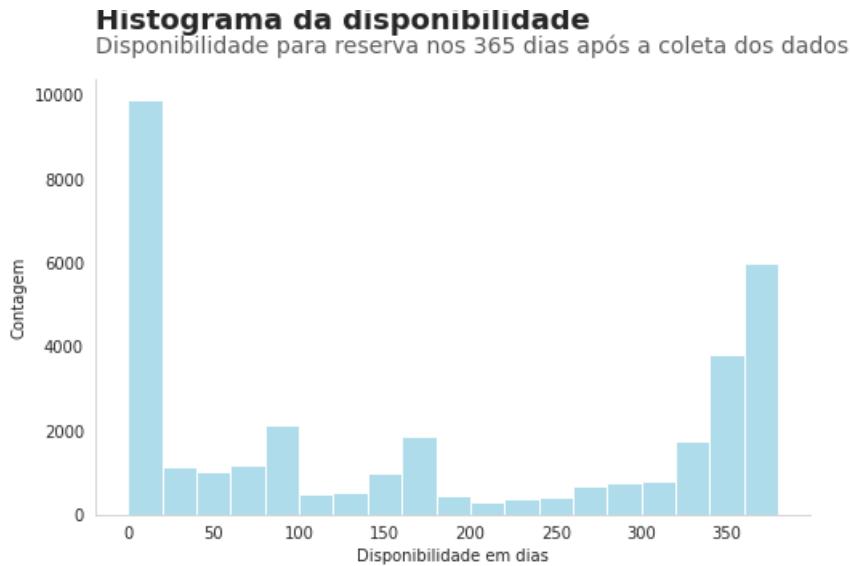
Preço (reais)	Número de hóspedes				Total
	Até 2	3-4	5-7	Mais de 7	
0-200	53%	35%	10%	2%	100%
201-500	26%	43%	26%	5%	100%
501-1000	13%	36%	38%	14%	100%
Mais de 1000	6%	23%	40%	31%	100%
Total	30%	37%	24%	9%	100%

Os anúncios no Airbnb podem receber avaliações dos hóspedes depois do período de estadia. A avaliação consiste em um comentário aberto e notas quanto à limpeza, comunicação, check-in, precisão, localização e valor. Inicialmente acreditávamos que as avaliações seriam um fator chave para explicar o preço, mas ao fazer uma análise exploratória percebemos que 41% dos anúncios não tem nenhuma avaliação e que quase 32% tem apenas entre um e cinco comentários.

Uma variável presente no banco de dados é o número de dias disponíveis para reserva dentre os próximos 356 dias. Existem fatores que podem deixar essa análise

menos confiável: um imóvel pode ter poucos dias disponíveis porque todos os outros estão reservados, ou pode ser porque o anfitrião bloqueou o calendário e não disponibilizou datas. Como pode ser observado na Figura 5, o atributo apresenta um padrão curioso, com vários picos na distribuição. Quase 25% dos anúncios tem zero dias disponíveis para reserva; enquanto 25% tem 350 dias ou mais, sendo que 9,2% tem todos os 365 dias vagos.

Figura 5 – Histograma da disponibilidade de reserva



Pela Tabela 3 observamos que de forma geral 40% dos espaços tem mais de 250 dias de disponibilidade. Para os imóveis com preço acima de quinhentos reais o percentual da disponibilidade igual a zero é maior do que nos outros imóveis.

Tabela 3 – Disponibilidade de reserva por faixas de preço

Preço (reais)	Disponibilidade de reserva nos próximos 365 dias				
	0	1-90	91-250	Mais de 250	Total
0-200	21%	22%	19%	39%	100%
201-500	20%	20%	19%	41%	100%
501-1000	32%	16%	11%	41%	100%
Mais de 1000	40%	11%	7%	43%	100%
Total	25%	19%	16%	40%	100%

Pelo Airbnb os hóspedes podem ver quais comodidades cada espaço oferece. Essas comodidades são marcadas pelo anfitrião e são atributos relativos ao espaço, como a presença de cozinha e estacionamento. Existem 172 tipos diferentes de comodidades, sendo que 91% dos espaços possuem cozinha e 89% possuem Wi-fi (Figura 6). Essas características apontam para a realidade de que os imóveis alugados são de fato casas completas; sendo que 67% dos espaços possuem até máquina de lavar roupa.

Dentre comodidades menos frequentes (Figura 7), temos itens como vista de frente ao mar, jacuzzi e jardim.

Figura 6 – Comodidades mais frequentes

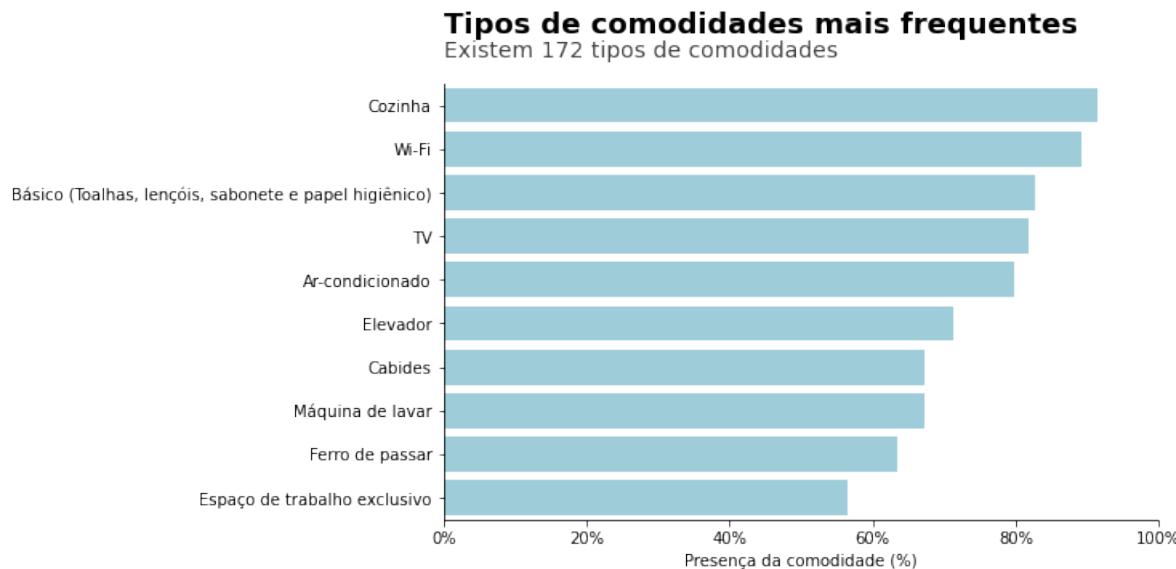
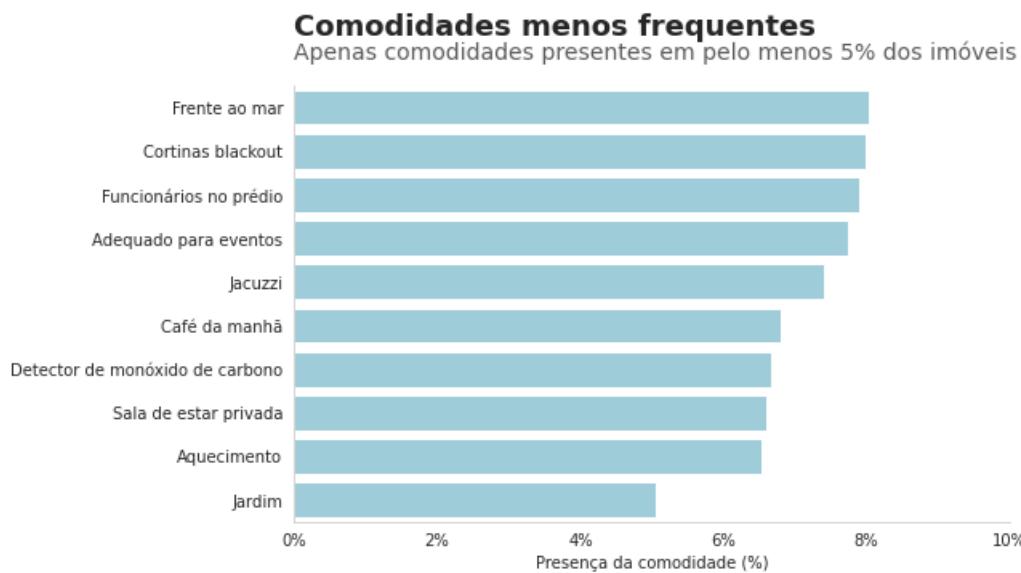
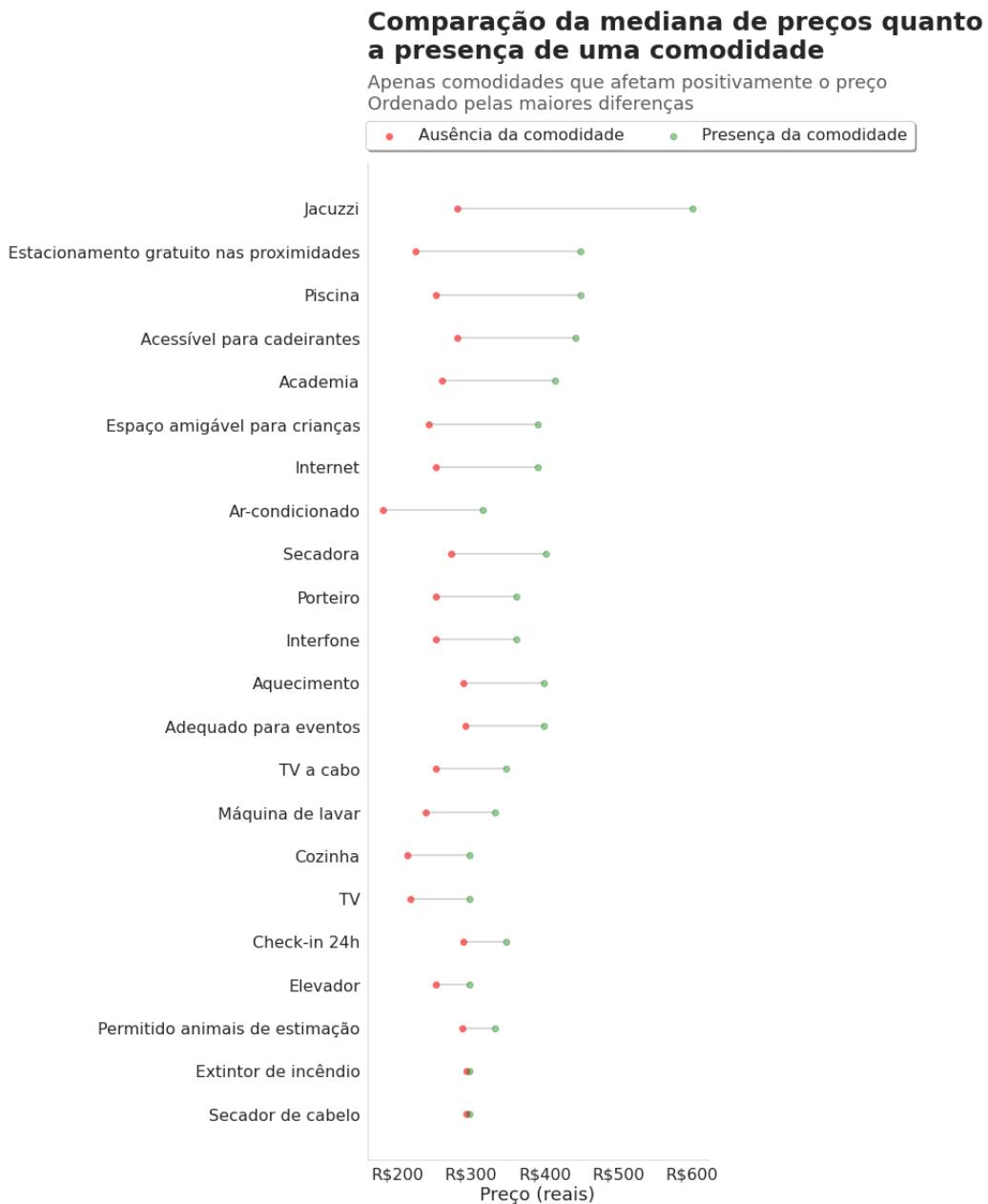


Figura 7 – Comodidades menos frequentes



Como um dos objetivos deste trabalho é identificar fatores que possam levar um imóvel a ter um aluguel mais caro, iremos detalhar a seguir uma análise mais completa em relação ao impacto das comodidades no preço, com uma comparação da mediana dos preços quando um imóvel tem uma comodidade ou não. Na Figura 8 apresentamos as comodidades com maior efeito de aumento na mediana do preço. Jacuzzi é o atributo com maior diferença nas medianas: 281 reais o valor da diárida sem jacuzzi e 601 reais com a presença. Estacionamento incluso é o segundo fator de maior diferença, seguido por piscina. O item ar-condicionado, que está presente em 80% dos imóveis, implicou em uma redução de 135 reais na mediana do preço quando está ausente.

Figura 8 – Efeito positivo de comodidades na mediana do preço



Na Figura 9 apresentamos as comodidades com os maiores efeitos negativos no preço mediano. Alguns dos resultados são contra intuitivos. Por exemplo, com a presença de água quente foi notada uma queda no preço mediano no valor de 181 reais. Uma hipótese é de que não é uma ausência verdadeira do item, mas sim a falta de marcação por parte do anfitrião. Por ser considerada uma comodidade básica, talvez anfitriões com imóveis repletos de itens não marquem a presença de água quente. Enquanto anfitriões de imóveis com poucas comodidades talvez marquem todos atributos, incluindo os mais simples. Uma variável que vai de acordo com o senso comum, é a queda do preço quando o estacionamento do imóvel é pago e fora da propriedade ou é na rua.

Para tentar entender o perfil dos espaços alugados, fizemos uma análise de corres-

Figura 9 – Efeito negativo de comodidades na mediana do preço



pondênciam com as comodidades oferecidas por cada imóvel. Dentre as regras de associação com confiança maior que 80% e com os maiores suportes, temos regras com as comodidades que são as mais frequentes: ar-condicionado, básico (tolhas, lençóis, sabonete e papel higiênico), cabides, ferro de passar, TV, Wi-Fi, cozinha.

Para tentar analisar outros perfis de espaço, selecionamos apenas regras com algumas comodidades específicas e retiramos as comodidades mais comuns. Primeiro, escolhemos regras que continham piscina. Identificamos que se um imóvel tem academia, ele tem uma probabilidade de 91% de também conter piscina; essa regra tem um suporte de quase 17%. Além de associações de piscina com as comodidades mais comuns, notamos também algumas regras associadas com estacionamento gratuito na propriedade. Em relação à comodidade

jacuzzi, a que mais causa uma diferença na mediana do preço, identificamos que em 82% das vezes que temos a jacuzzi também temos elevador, contrariando uma hipótese inicial que as hidromassagens seriam exclusivamente em casa e apontando para possíveis apartamentos com coberturas com varandas. Dos imóveis com jacuzzi, identificamos também que 79% tem estacionamento gratuito, outro fator que contribui para a valorização do espaço.

Em relação à comodidade de acessibilidade para cadeirantes, identificamos uma regra coerente: se um espaço é acessível, ele tem 91% de chance de ter elevador. Outras regras coerentes identificadas foram as associações entre porteiro, elevador e interfone; e entre os itens de cozinha, como fogão, geladeira e microondas.

Em relação aos novos atributos criados, conseguimos identificar o sexo de 93% dos anfitriões a partir do nome. Dentre os identificados, 52% são do sexo feminino e 48% do sexo masculino. Já para a variável que avalia o idioma das avaliações, temos que comentários em português são, em mediana, 57% dos textos de cada imóvel.

3.3 Localização dos imóveis

Além de ajustar modelos de regressão para a precificação, temos como objetivo identificar padrões ou fatores que possam levar um imóvel a ter um aluguel mais elevado. Nesta Seção, iremos investigar se a localização é um dos fatores que afeta o preço. Na Figura 10 temos um *heatmap* da quantidade de Airbnbs na cidade do Rio de Janeiro. É possível perceber pontos fortes na região de Copacabana, Ipanema e Leblon. A única região fora da área das praias que tem uma densidade de Airbnbs é perto da Avenida Embaixador Abelardo Bueno, que é uma avenida que atravessa a divisa entre os bairros de Jacarepaguá e Barra da Tijuca. Além disso, a região é bem próxima à Cidade do Rock, local onde acontece o evento Rock in Rio, e próxima ao local de vários eventos da Olimpíada de 2016, incluindo a vila olímpica.

Na Figura 11 temos o *heatmap* dos imóveis considerando o preço. Comparando com o mapa anterior, percebemos que a região central, apesar de ter uma concentração razoavelmente alta de imóveis, não reflete em preços tão altos. Na Figura 12 temos o mapa com um zoom na região mais nobre, de Copacabana; e na Figura 13 uma visão da região do Centro, que apesar de ter uma quantidade alta de imóveis e estar em uma região perto de pontos turísticos (Figura 14), não é uma área de alto custo para o aluguel.

Figura 10 – Heatmap dos imóveis Airbnb na cidade do Rio de Janeiro

Heatmap dos imóveis Airbnb na cidade do Rio de Janeiro

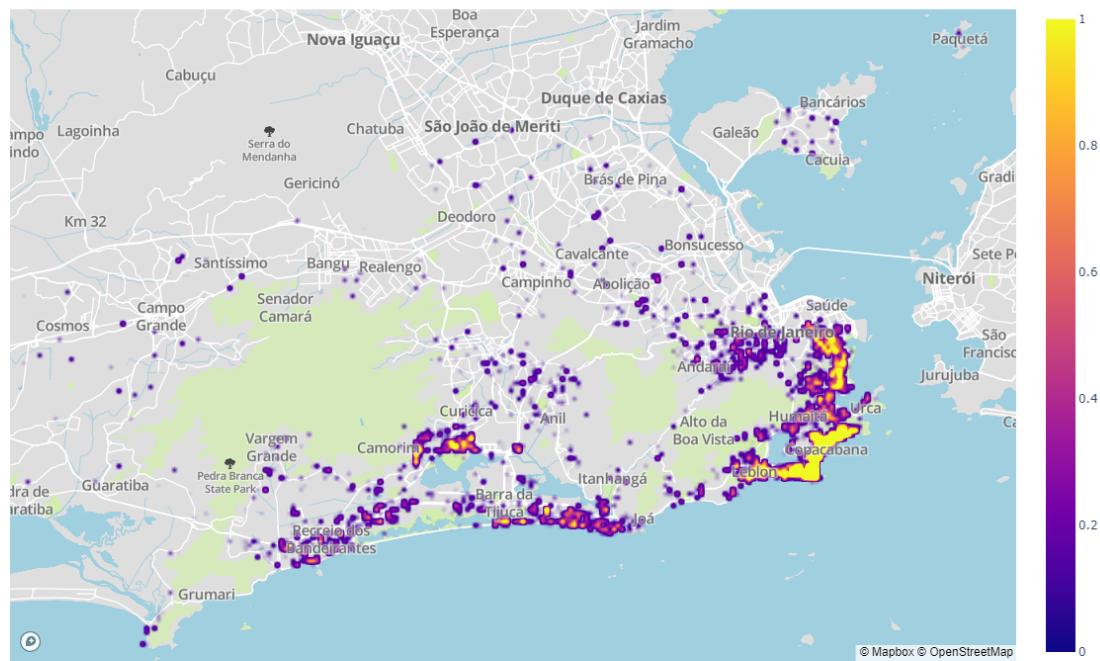


Figura 11 – Heatmap dos imóveis Airbnb para aluguel na cidade do Rio de Janeiro por preço

Heatmap dos imóveis Airbnb para aluguel por preço

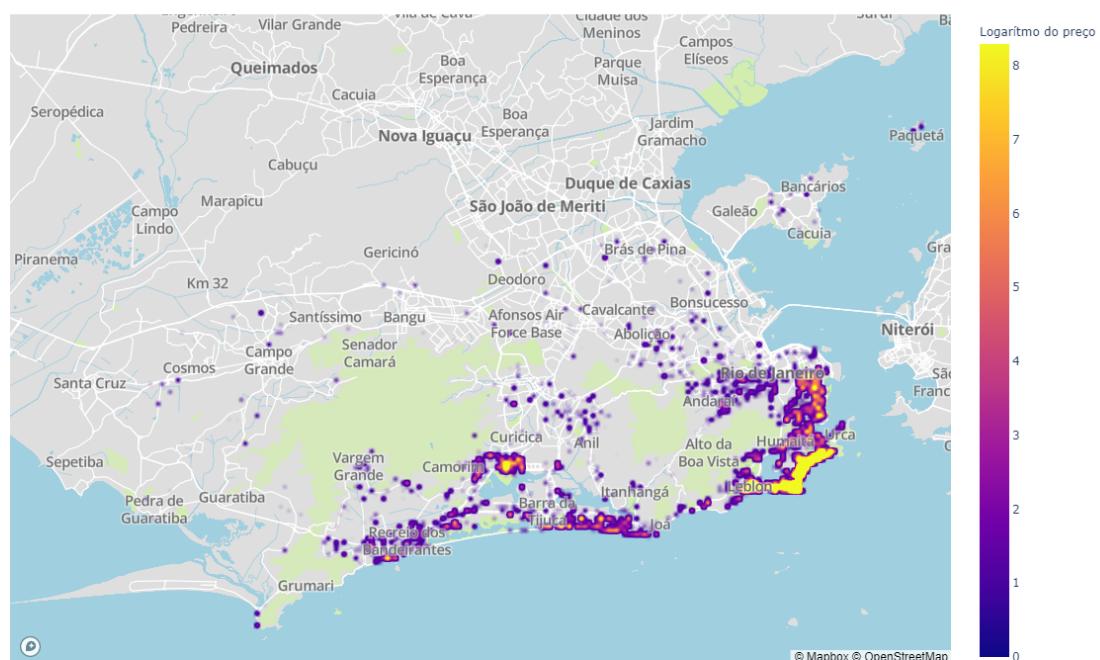


Figura 12 – Heatmap dos imóveis Airbnb para aluguel por preço na região do Leblon, Ipanema e Copacabana

Heapmap dos imóveis Airbnb para aluguel por preço - Leblon, Ipanema e Copacabana

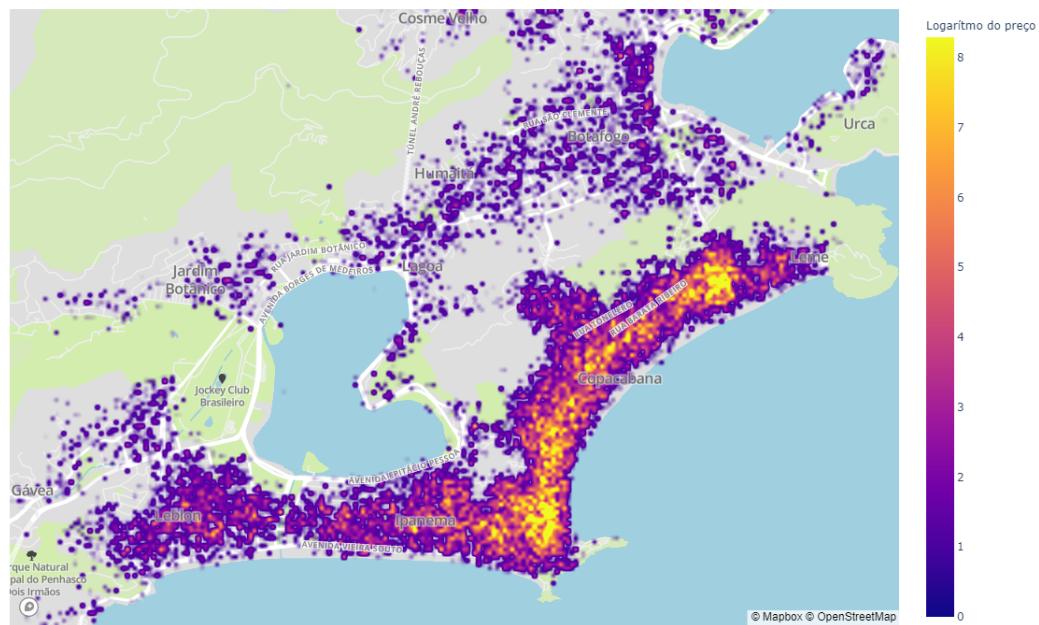


Figura 13 – Heatmap dos imóveis Airbnb para aluguel por preço na região do Centro

Heapmap dos imóveis Airbnb para aluguel por preço - Centro

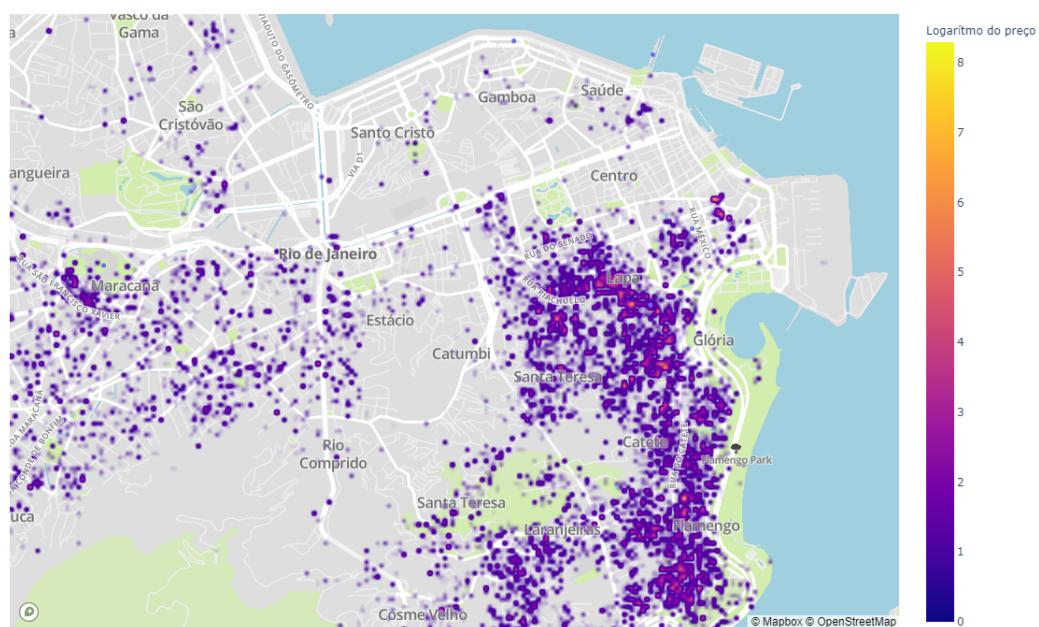
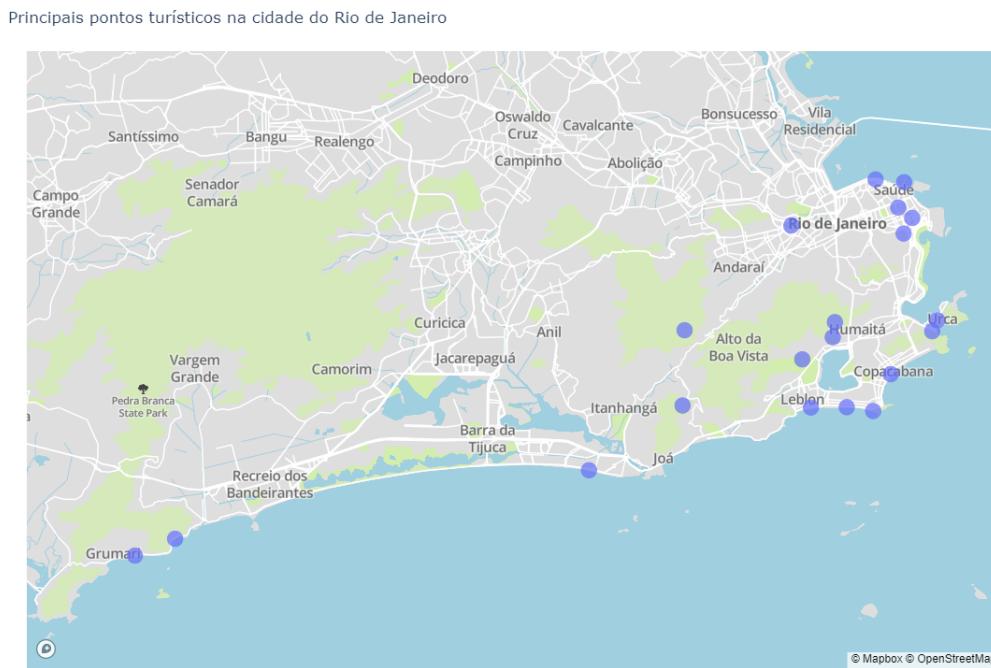


Figura 14 – Principais pontos turísticos na cidade do Rio de Janeiro



3.4 Análises dos atributos em formato texto

Vários dos atributos referentes aos espaços de aluguel são textos, incluindo o nome do anúncio e as avaliações que os hóspedes podem escrever.

Para os nomes anunciados, primeiro fizemos uma detecção de idioma. Observamos que 51% dos nomes foram identificados como do idioma português e 28% em inglês. Retiramos as *stopwords* dos nomes, que são palavras consideradas irrelevantes, e criamos então uma nuvem de palavras para cada idioma. Na Figura 15 temos a nuvem de anúncios em português, onde palavras como “Rio Janeiro”, “quarto”, “Apartamento”, “Copacabana”, “Ipanema” e “Barra Tijuca” se destacam. Na nuvem de nomes em inglês, Figura 16, as principais palavras foram parecidas, sendo que “Copacabana” e “Ipanema” tiveram um grande destaque. Interessante notar que tanto na nuvem em português quanto na nuvem em inglês a palavra “Olimpíada” aparece, indicando possíveis anúncios criados na época do evento esportivo, que aconteceu em 2016, e não atualizados desde então.

Em relação às avaliações, 44% dos imóveis não tem nenhum comentário. Dentre os espaços que receberam notas, a grande maioria apresenta notas muito altas, sendo a mediana um valor de 98 em 100. Mais uma vez fizemos a detecção dos idiomas e construímos nuvens separadas. Na nuvem dos comentários em português (Figura 17) palavras referentes à localização apareceram com grande destaque. Para tentar entender as avaliações ruins, filtramos todos os comentários com notas abaixo de 70%, o que resultou em apenas 472 comentários. Pela nuvem apresentada na Figura 18 notamos palavras como “problema”,

Figura 15 – Nuvem de palavra dos nomes de anúncio em português dos imóveis

Nuvem de palavras dos nomes de anúncio dos imóveis

Nomes em português

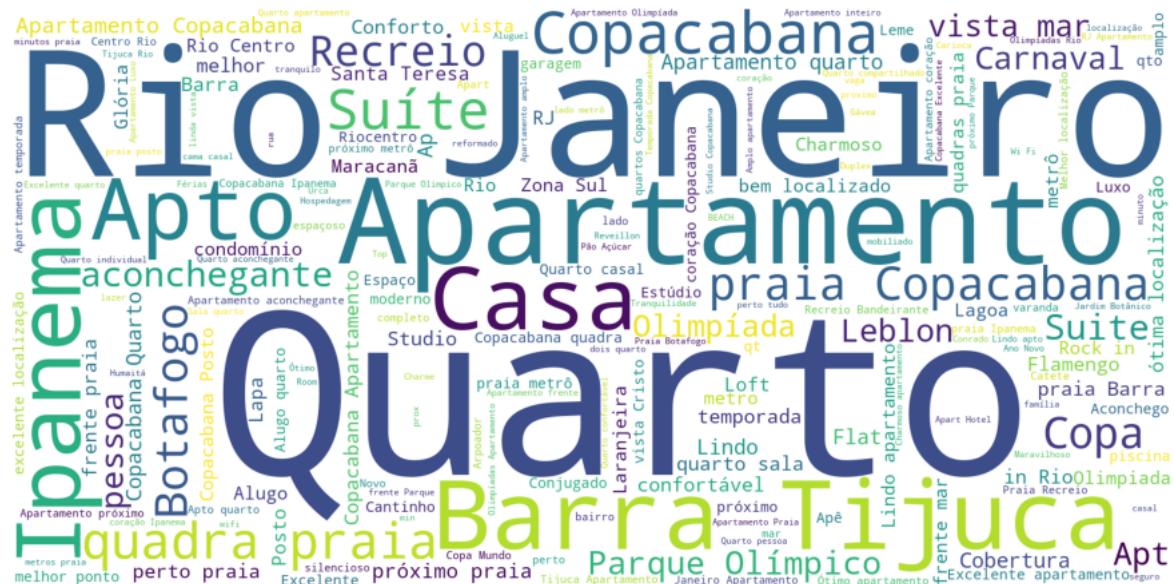


Figura 16 – Nuvem de palavra dos nomes de anúncio em inglês dos imóveis

Nuvem de palavras dos nomes de anúncio dos imóveis

Nomes em inglês



“limpeza”, “espaço sujo”, “anfitrião” e “ar condicionado”.

A nuvem de comentários em inglês, Figura 19, apresenta palavras chaves parecidas com a nuvem em português, como palavras referentes à localização e anfitrião.

Figura 17 – Nuvem de palavra para comentários em Português

Nuvem de palavras dos comentários

Comentários em português



Figura 18 – Nuvem de palavra para comentários de avaliações ruins em Português

Nuvem de palavras dos comentários

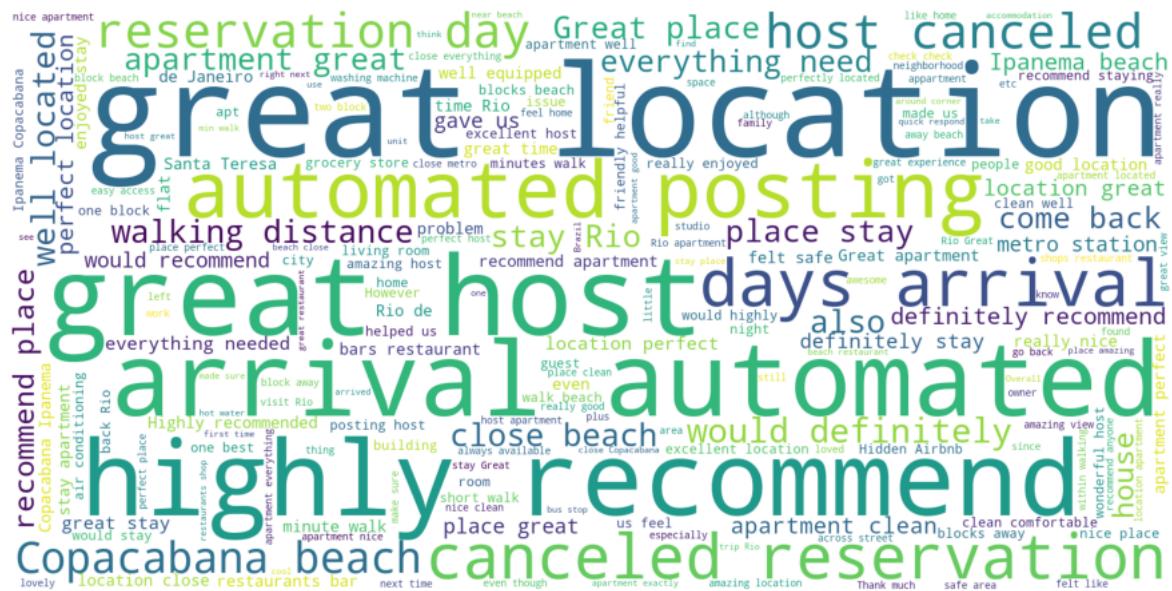
Comentários em português de imóveis com avaliação menor do que 70% - Apenas 472 comentários



Figura 19 – Nuvem de palavra para comentários em Inglês

Nuvem de palavras dos comentários

Comentários em inglês



3.5 Resultados da precificação

Realizado o pré-processamento da base de dados e concluído um entendimento inicial do problemas através das análises descritivas, agora começamos a etapa de modelagem. Selecionamos quatro modelos: o modelo de regressão linear, para seguir como base de comparação; regressão *Ridge*, já que ele penaliza o número de parâmetros em excesso; e os modelos floresta aleatória e *gradient boosting*, que são alternativas mais complexas.

Inicialmente, para os modelos regressão *Ridge*, floresta aleatória e *gradient boosting* não foram ajustados hiperparâmetros, sendo utilizado as configurações padrões e automáticas que a biblioteca *scikit-learn* fornece. Os resultados estão presentes da Tabela 4; percebemos que apenas o R^2 na base de treino do modelo floresta aleatória é muito bom, de 0,9. No entanto, essa performance não se estende à base de teste, onde se observa um R^2 de 0,23. Essa diferença de performance entre as partições dos dados é um claro exemplo de *overfitting*.

Tabela 4 – Resultados - Variáveis originais

Algorítimo	RMSE - Treino	RMSE - Teste	R^2 - Treino	R^2 - Teste
Regressão linear	1.443,21	1.577,35	0,24	0,17
Regressão <i>Ridge</i>	1.443,28	1.577,17	0,24	0,17
Floresta aleatória	527,88	1.514,74	0,90	0,23
<i>Gradient boosting</i>	1.170,47	1.475,01	0,50	0,27

Como existem muitos atributos com distribuição assimétrica, optamos por tentar usar transformações logarítmicas para melhorar o ajuste, inclusive na variável resposta.

que possui diária máxima de quase 42 mil reais. Como pode ser observado na Tabela 5, as transformações fizeram uma grande diferença. Nos modelos de regressão linear e *Ridge*, observamos R^2 entorno de 0,63 e 0,62, nas bases de treino e teste, se comparados aos valores anteriores de 0,24 e 0,17. O modelo de floresta aleatória apresentou uma performance ainda melhor na base de treino e, apesar de ainda ter indícios de *overfitting*, o R^2 de teste melhorou consideravelmente; de 0,23 para 0,67. O modelo *gradient boosting* tem um ajuste em teste parecido com o floresta aleatória, com o R^2 em 0,66, mas sem problemas de *overfitting*.

Tabela 5 – Resultados - Variáveis com transformações

Algorítimo	RMSE - Treino	RMSE - Teste	R^2 - Treino	R^2 - Teste
Regressão linear	0,63	0,62	0,63	0,62
Regressão <i>Ridge</i>	0,63	0,62	0,63	0,63
Floresta aleatória	0,22	0,58	0,95	0,68
<i>Gradient boosting</i>	0,58	0,59	0,68	0,66

Como os modelos floresta aleatória e *gradient boosting* foram os que apresentaram melhor desempenho, realizamos um ajuste dos seus hiperparâmetros, conforme explicado na subseção 2.3.5, para tentar melhorar a performance. No caso do modelo de floresta aleatória, usamos parâmetros para tentar limitar o *overfitting*. Como observado na Tabela 6, conseguimos diminuir o problema de *overfitting* do modelo de floresta aleatória, mas o R^2 na base de teste acabou diminuindo também. Para o modelo *gradient boosting*, não conseguimos uma melhora significativa no ajuste.

Tabela 6 – Resultados - Modelos com ajuste de hiperparâmetros

Algorítimo	RMSE - Treino	RMSE - Teste	R^2 - Treino	R^2 - Teste
F.Aleatória - Padrão	0,22	0,58	0,95	0,67
F.Aleatória - Ajustada	0,61	0,63	0,65	0,61
<i>G.Bosting</i> - Padrão	0,58	0,59	0,68	0,66
<i>G.Bosting</i> - Ajustada	0,57	0,59	0,69	0,66

Para os ajustes mencionados anteriormente, usamos todas as 188 variáveis. Como esse é um número de atributos muito elevado, tentamos fazer uma seleção das variáveis a partir do modelo *Ridge* mostrado na Tabela 5, já que ele penaliza o número excessivo de atributos. Duas seleções de variáveis foram consideradas, uma mais moderada e outra um pouco mais rígida. A primeira, excluiu variáveis com o valor de coeficiente absoluto no modelo *Ridge* menor do que 0,03 e implicou na redução para 125 colunas. A segunda seleção excluiu variáveis com o coeficiente absoluto menor do que 0,05, resultando no uso de 53% das variáveis originais. Como pode ser visto na Tabela 7, a redução moderada das variáveis não diminuiu muito o poder preditivo dos modelos.

Como os dados são muito heterogêneos, tentamos um último ajuste onde ajustamos um modelo para alguns tipos de acomodação diferente (Tabela 8). Os melhores resultados

Tabela 7 – Resultados - Modelos após seleção de variáveis

Algorítimo	RMSE - Treino	RMSE - Teste	R^2 Treino	R^2 Teste	Qtde variáveis
F.Aleatória	0,22	0,58	0,95	0,68	188 (100%)
	0,23	0,59	0,95	0,66	125 (66%)
	0,23	0,61	0,95	0,63	99 (53%)
<i>G.Boosting</i>	0,58	0,59	0,68	0,66	188 (100%)
	0,60	0,61	0,65	0,64	125 (66%)
	0,62	0,63	0,63	0,62	99 (53%)

Tabela 8 – Resultados - Modelos diferentes para cada tipo de acomodação

Acomodação	Algorítimo	RMSE - Treino	RMSE - Teste	R^2 Treino	R^2 Teste
Espaço int.	R.Linear	0,60	0,60	0,62	0,61
	R. <i>Ridge</i>	0,60	0,60	0,62	0,61
	F.Aleatória	0,21	0,57	0,95	0,64
	<i>G.Boosting</i>	0,55	0,56	0,68	0,65
Quarto int.	R.Linear	0,62	0,62	0,42	0,42
	R. <i>Ridge</i>	0,62	0,62	0,42	0,42
	F.Aleatória	0,23	0,59	0,92	0,48
	<i>G.Boosting</i>	0,55	0,59	0,55	0,49
Quarto comp.	R.Linear	0,57	0,97	0,58	0,06
	R. <i>Ridge</i>	0,58	0,92	0,56	0,16
	F.Aleatória	0,27	0,85	0,90	0,28
	<i>G.Boosting</i>	0,39	0,82	0,80	0,34

são com a partição de dados de espaços do tipo espaço inteiro, com um R^2 na base de teste de 0,65. Já os tipos quarto inteiro e quarto compartilhado tiveram um desempenho pior, um R^2 de 0,49 e 0,34 respectivamente nas bases de teste.

4 DISCUSSÕES E CONCLUSÕES

Nesta seção, apresentamos mais detalhes sobre alguns dos modelos que trouxeram melhores resultados na tarefa de predição do preço.

O modelo que apresentou um melhor desempenho, tanto pelo RMSE quanto pelo R^2 , foi o *gradient boosting*. O ajuste dos hiperparâmetros do modelo não trouxe ganhos significativos, sendo observado um R^2 de 0,96 em treino e 0,66 em teste. O modelo de floresta aleatória tem um desempenho que tende muito ao *overfitting*, inicialmente apresentando R^2 na base de treino de 0,95 e 0,67 em teste. Para esse modelo, quando fazemos ajuste dos hiperparâmetros e tentamos controlar o *overfitting*, o desempenho final acaba sendo pior que o do *boosting*, com R^2 de 0,65 em treino e 0,61 em teste.

Para o modelo considerado melhor, o *gradient boosting* com ajuste dos hiperparâmetros, identificamos as variáveis de maior importância para as criações dos nós das árvores. Como pode ser visto na Figura 20, a quantidade de camas de um imóvel é um fator primordial. Em seguida, temos a quantidade de banheiros e o número de avaliações, sendo que para o último, interessante lembrar que 41% dos anúncios não tem nenhuma avaliação e quase 32% tem apenas entre um e cinco comentários. O tipo de comodidade água quente, que já havíamos identificado com uma tendência contraintuitiva de queda no preço com a sua presença (Figura 9), foi a décima quinta variável de maior importância para o modelo. Dentre as variáveis criadas, a mais importante foi a distância até o ponto turístico da praia da Barra da Tijuca.

Das 188 variáveis consideradas no modelo, 115 tiveram a importância estimada muito baixa, reforçando a necessidade de uma seleção de variáveis. Como descrito na seção anterior na Tabela 7, fizemos uma seleção de variáveis a partir das variáveis que o modelo *Ridge* atribuía valores muito pequenos. Para o *gradient boosting*, o modelo com todas as 188 variáveis obtemos um R^2 na base de teste de 0,66, considerando 66% das variáveis o R^2 é de 0,64, e com 53% dos atributos observamos um R^2 de 0,62. Apesar da queda na métrica de desempenho, é uma queda moderada para um corte tão grande de variáveis, que tenta diminuir a complexidade do modelo.

Dentre as variáveis criadas, algumas tiveram importância para o modelo. Para as de reconhecimento do idioma das avaliações, a mais importante foi o percentual de avaliações em português, seguido de inglês. Para a variável de distância em quilômetros até os pontos turísticos, várias tiveram alguma importância identificada. Por ordem de relevância, temos praia da Barra da Tijuca, Pedra Bonita, praia do Leblon e praia de Copacabana. Já para a hipótese de que o sexo do anfitrião poderia afetar o preço, não conseguimos encontrar evidências, já que a variável não foi importante para o modelo.

Figura 20 – As vinte variáveis mais importantes



Para os ajustes apresentados, enfrentamos algumas dificuldades. A principal delas foi a complexidade e heterogeneidade dos dados, e a dificuldade de validar os vários casos extremos. Como os dados são obtidos via *web scraping* no site do Airbnb, não conseguimos ter a garantia que os anúncios estão ativos. A seguir listaremos alguns indícios dessa complexidade.

Como vimos nas nuvens de palavras (Figura 15 e Figura 16), a palavra “olimpíada” aparece nos nomes de anúncios em português e em inglês. O evento esportivo aconteceu em 2016 e é bem improvável que um espaço com anúncio ativo ainda use essa palavra como propaganda. Acreditamos que por ser um evento de atração mundial, pessoas alugaram temporariamente suas próprias casas por um preço acima do usual.

Outra questão que mostra a complexidade dos dados pode ser vista através da variável de dias disponíveis para reserva dentre os próximos 365 dias. Quase 25% dos anúncios tem zero dias disponíveis para reserva. Aqui temos duas teorias: o imóvel realmente tem todos os 365 dias reservados, o que parece improvável, ou o anfitrião bloqueou parte das datas do calendário. Considerando o segundo caso, ainda temos outros dois cenários. O anfitrião pode ter bloqueado os 365 dias e o anúncio não estar mais ativo, ou o anfitrião pode ter bloqueado grande parte do calendário, deixando apenas alguns dias disponíveis e todos esses foram reservados. Em situação oposta, 25% dos espaços tem 350 dias vagos ou mais, sendo que 9,2% tem todos os 365 dias vagos. Para este cenário, é possível que existam anúncio não ativos, já que deve ser improvável um anúncio não ter reserva para nenhum dia.

Uma variável para ajudar a entender se o espaço está ativo ou não é a última atualização feita pelo anfitrião no calendário. Em 47% dos anúncios o anfitrião atualizou o calendário nas últimas sete semanas, 16% entre dois e seis meses, e 37% atualizaram a mais de seis meses. Existe um imóvel com data de atualização de 95 meses, o que corresponde a quase oito anos. Não podemos afirmar com certeza, mas é bem possível que esses anúncios não estejam mais abertos para receber hóspedes, apesar de estarem presentes no site.

O número mínimo de diárias no imóvel é outro exemplo de variável com valores extremos. 51% dos anúncios pedem uma ou duas diárias mínimas, mas existe um caso de estadia mínima de 1123 dias, que é pouco mais de três anos. Cinco espaços possuem estadia mínima de mil ou mais dias. Ao analisar esses casos individualmente, notamos ao tentar abrir a url que os anúncios não existem mais. Em relação aos preços, eles variaram, sendo o menor no valor de 83 reais a diária e o máximo de quase cinco mil reais.

Para a nossa variável resposta preço existem muito valores extremos que também não temos como validar se são realmente de imóveis ativos. Dois porcento dos imóveis tem valores de diárias acima de quatro mil reais. Abrimos anúncios aleatoriamente e em alguns poucos casos o preço parece coerente por se tratar de um imóvel de luxo, mas na maioria dos anúncios investigados o preço parece não se justificar.

De forma geral, é complicado fazer a validação dos anúncios. Tentamos fazer o ajuste do modelo de precificação com algumas considerações, por exemplo apenas para preços abaixo de cinco mil reais e também para os com disponibilidade de dias diferente de zero e de 365 dias; mas não obtemos melhora.

Como entendemos que os dados e as relações das variáveis são complexas, tentamos retirar as dezesseis variáveis que fizemos imputação e refazer a modelagem, mas não foi observada diferença nos critérios de seleção de modelo que estamos considerando. A maioria das variáveis com valores faltantes são relacionadas às avaliações. Como em 44% dos imóveis não temos avaliações, não existem notas relacionadas à limpeza, comunicação, check-in, precisão, localização e valor. Acreditamos que essas avaliações são muito importantes para avaliar o preço de um espaço, e ao imputar a mediana podemos estar sendo simplista demais.

4.1 Considerações finais e trabalhos futuros

O Airbnb é um dos tipos de serviços da chamada economia compartilhada. Através do site pessoas podem obter um dinheiro extra ao alugar recursos que estão subutilizados, como um quarto parado em casa. Percebemos que apesar do Airbnb ter surgido com essa proposta, 72% dos espaços no Rio de Janeiro são do tipo espaço inteiro. Esse fato sugere que pessoas são proprietárias de imóveis além das suas próprias casas, deixando um imóvel exclusivamente para aluguel de Airbnb, se tornando uma fonte de renda adicional.

Esse trabalho contribuiu para o entendimento da precificação de Airbnb, utilizando técnicas de aprendizado de máquina. De acordo com os objetivos definidos na Seção ??, estudamos os modelos de predição comumente empregados e comparamos a performance de diferentes algorítimos. Enfrentamos muitos problemas quanto à complexidade e heterogeneidade dos dados. O nosso melhor modelo obteve um R^2 de 0,66 na base de treino, enquanto o estudo de [Kalehbasti, Nikolenko e Rezaei \(2019\)](#) apresentou um valor de 0,69. O trabalho de [Luo, Zhou e Zhou \(2019\)](#) mostra que um modelo treinado em um conjunto de dados combinado de duas cidades, ao invés de conjuntos de dados individuais, é mais generalizável para prever o preço em uma nova cidade; dessa forma, conseguiram um R^2 maior, de 0,77.

Em concordância com o objetivo de identificar fatores que levam um imóvel a ter um aluguel mais elevado, entendemos que a localização é um fator primordial para a valorização do imóvel, e algumas comodidades são extremamente comuns entre todos os espaços disponíveis, como cozinha, Wi-fi, televisão e ar-condicionado. A presença de comodidades menos comuns, como jacuzzi, piscina e estacionamento estão entre os fatores que mais causam uma diferença de acréscimo no preço.

Para pontos de melhoria e trabalhos futuros, acreditamos que estudos envolvendo análise espacial e processamento de linguagem natural poderiam melhorar o entendimento do problema e talvez levar a um ganho na precificação, já que muitas das variáveis disponíveis são textuais.

Como enfrentamos uma dificuldade de validar se os anúncios estão ativos e provavelmente esses casos estão dificultando o nosso ajuste, outro ponto de melhoria seria verificar se cada URL segue válido atualmente. Os dados considerados para a pesquisa foram coletados no final do mês de janeiro de 2020, quando todos estavam vigentes, mas um anúncio não funcionar atualmente é um indício de que ele poderia já não estar com informações atualizadas e corretas.

A localização do espaço parece ser um dos pontos mais importantes para a sua valorização. Nesse sentido, seria interessante evoluir as análises de distância, envolvendo análises espaciais, por exemplo. Um ponto inicial seria não apenas considerar a distância em linha reta até os pontos turísticos, mas considerar o tempo de deslocamento usando, por exemplo, a API do Google Maps.

Outro plano de evolução do trabalho seria traçar perfis dos imóveis alugados no Airbnb. Para traçar esses perfis teríamos um entendimento melhor de quais comodidades um imóvel tem, desde o nível mais simples até o luxuoso. A partir desses perfis, diferenciando também por tipo de acomodação, considerar a precificação separadamente.

Finalmente, outra possibilidade de estudo é seguir o resultado proposto por [Kalehbasti, Nikolenko e Rezaei \(2019\)](#), que considera um treino inicial de uma rede neural em

um conjunto de dados combinado de mais cidades.

REFERÊNCIAS

- AGRAWAL, R.; IMIELIŃSKI, T.; SWAMI, A. Mining association rules between sets of items in large databases. In: **Proceedings of the 1993 ACM SIGMOD international conference on Management of data**. [S.l.: s.n.], 1993. p. 207–216.
- AIRBNB. **Conheça o Airbnb?** 2020. Disponível em: <<https://www.airbnb.com.br/d/howairbnbwoks>>.
- _____. **Primeiros passos para começar a usar o Airbnb.** 2020. Disponível em: <<https://www.airbnb.com.br/resources/hosting-homes/a/the-essentials-get-started-on-airbnb-1>>.
- BARRON, K.; KUNG, E.; PROSERPIO, D. The effect of home-sharing on house prices and rents: Evidence from airbnb. **Available at SSRN 3006832**, 2018.
- BORGELT, C.; KRUSE, R. Induction of association rules: Apriori implementation. In: SPRINGER. **Compstat**. [S.l.], 2002. p. 395–400.
- BRASIL, C. **Avanços e dúvidas no caminho do compartilhamento.** 2016. Disponível em: <<https://www.pwc.com.br/pt/publicacoes/revista-ceo/assets/2016/PwC-CEO-BRASIL-31.pdf>>.
- DESBOULETS, L. D. D. A Review on Variable Selection in Regression Analysis. **Econometrics**, v. 6, n. 4, p. 1–27, November 2018. Disponível em: <<https://ideas.repec.org/a/gam/jecnmx/v6y2018i4p45-d185046.html>>.
- EAGLESHAM, K. G. J. **Airbnb Paying More Than 10Financing Announced Monday.** 2020. Disponível em: <<https://www.wsj.com/articles/airbnb-paying-more-than-10-interest-on-1-billion-financing-announced-monday-11586297484>>.
- EDELMAN, B.; LUCA, M.; SVIRSKY, D. Racial discrimination in the sharing economy: Evidence from a field experiment. **American Economic Journal: Applied Economics**, v. 9, n. 2, p. 1–22, 2017.
- EDELMAN, B. G.; LUCA, M. Digital discrimination: The case of airbnb. com. **Harvard Business School NOM Unit Working Paper**, n. 14-054, 2014.
- FRIEDMAN, J. H. Greedy function approximation: A gradient boosting machine. **Ann. Statist.**, The Institute of Mathematical Statistics, v. 29, n. 5, p. 1189–1232, 10 2001. Disponível em: <<https://doi.org/10.1214/aos/1013203451>>.
- GARRETT, F. **O que é AirBnb?** 2017. Disponível em: <<https://www.techtudo.com.br/dicas-e-tutoriais/noticia/2017/01/o-que-e-airbnb.html>>.
- JANEIRO, P. da cidade do Rio de. **Melhor carnaval de todos os tempos no Rio: mais de 10 milhões de foliões e alto índice de aprovação por turistas.** 2020. Disponível em: <<https://prefeitura.rio/rio-acontece/melhor-carnaval-de-todos-os-tempos-no-rio-mais-de-10-milhoes-de-folioes-e-alto-indice-de-aprovacao-#:~:text=O%20Carnaval%20Rio%202020%20foi,bilh%C3%B5es%20em%20movimenta%C3%A7%C3%A3o%20econ%C3%B4mica%20e>>.

KALEHBASTI, P. R.; NIKOLENKO, L.; REZAEI, H. Airbnb price prediction using machine learning and sentiment analysis. **arXiv preprint arXiv:1907.12665**, 2019.

LEARN scikit. **1.1. Linear Models**. 2020. Disponível em: <https://scikit-learn.org/stable/modules/linear_model.html>.

_____. **3.1. Cross-validation: evaluating estimator performance**. 2020. Disponível em: <https://scikit-learn.org/stable/modules/cross_validation.html#cross-validation>.

_____. **3.2.4.3.2. sklearn.ensemble.RandomForestRegressor**. 2020. Disponível em: <<https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html#sklearn.ensemble.RandomForestRegressor>>.

_____. **3.2.4.3.6. sklearn.ensemble.GradientBoostingRegressor**. 2020. Disponível em: <<https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html>>.

_____. **6.4. Imputation of missing values**. 2020. Disponível em: <<https://scikit-learn.org/stable/modules/impute.html>>.

LUO, Y.; ZHOU, X.; ZHOU, Y. Predicting airbnb listing price across different cities. 2019.

MEIRELES, F. **genderBR: predizendo sexo a partir de nomes próprios**. 2017. Disponível em: <<https://fmeireles.com/blog/rstats/genderbr-predizer-sexo/>>.

PAULO, F. de S. **Brasil confirma primeiro caso do novo coronavírus**. 2020. Disponível em: <<https://www1.folha.uol.com.br/equilibrioesaude/2020/02/brasil-confirma-primeiro-caso-do-novo-coronavirus.shtml>>.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

PRICEWATERHOUSECOOPERS). **The sharing economy - Consumer intelligence series**. 2015. Disponível em: <https://www.pwc.fr/fr/assets/files/pdf/2015/05/pwc_etude_sharing_economy.pdf>.

TEAM, T. **As A Rare Profitable Unicorn, Airbnb Appears To Be Worth At Least \$38 Billion**. 2018. Disponível em: <<https://www.forbes.com/sites/greatspeculations/2018/05/11/as-a-rare-profitable-unicorn-airbnb-appears-to-be-worth-at-least-38-billion/?sh=4acab552741e>>.

TEUBNER, T.; HAWLITSCHEK, F.; DANN, D. Price determinants on airbnb: How reputation pays off in the sharing economy. **Journal of Self-Governance & Management Economics**, v. 5, n. 4, 2017.

Anexos

ANEXO A – LISTA DE VARIÁVEIS UTILIZADAS

Tabela 9 – Tabela com todas as variáveis utilizadas

Tema	ID	Descrição da variável	Tipo	Valores faltante
Anfitrião	1	Tempo de resposta do anfitrião	Categórico (4 categorias)	34,3%
Anfitrião	2	Nota 0-1 sobre a responsividade do anfitrião	Numérico	34,3%
Anfitrião	3	O anfitrião é "superhost"?	Boleano	0,1%
Anfitrião	4	Quantidade de anúncios do anfitrião	Numérico	0,1%
Anfitrião	5	Identidades verificadas pelo anfitrião	Lista (18 categorias)	0,0%
Anfitrião	6	Anfitrião tem foto?	Boleano	0,1%
Anfitrião	7	Identidade do anfitrião verificada?	Boleano	0,1%
Anfitrião	8	Número de anúncios do anfitrião	Numérico	0,0%
Anfitrião	9	Número de anúncios do anfitrião do tipo casa inteira	Numérico	0,0%
Anfitrião	10	Número de anúncios do anfitrião do tipo quarto privado	Numérico	0,0%
Anfitrião	11	Número de anúncios do anfitrião do tipo quarto compartilhado	Numérico	0,0%
Anfitrião	12	Sexo do anfitrião identificado a partir do nome	Categórico (2 categorias)	0,0%
Disponibilidade, requisitos e preços	13	Preço da diária	Numérico	0,0%
Disponibilidade, requisitos e preços	14	Preço do depósito de segurança	Numérico	43,3%
Disponibilidade, requisitos e preços	15	Preço da taxa de limpeza	Numérico	31,4%
Disponibilidade, requisitos e preços	16	Custo de hóspede adicional	Numérico	0,0%
Disponibilidade, requisitos e preços	17	Número mínimo de diárias	Numérico	0,0%

Disponibilidade, requisitos e preços	18	Número máximo de diárias	Numérico	0,0%
Disponibilidade, requisitos e preços	19	Número mínimo do mínimo de diárias	Numérico	0,0%
Disponibilidade, requisitos e preços	20	Número máximo do mínimo de diárias	Numérico	0,0%
Disponibilidade, requisitos e preços	21	Número mínimo do máximo de diárias	Numérico	0,0%
Disponibilidade, requisitos e preços	22	Número máximo do máximo de diárias	Numérico	0,0%
Disponibilidade, requisitos e preços	23	Média do número mínimo de diárias	Numérico	0,0%
Disponibilidade, requisitos e preços	24	Média do número máximo de diárias	Numérico	0,0%
Disponibilidade, requisitos e preços	25	Dias disponíveis dentre os próximos 30	Numérico	0,0%
Disponibilidade, requisitos e preços	26	Dias disponíveis dentre os próximos 60	Numérico	0,0%
Disponibilidade, requisitos e preços	27	Dias disponíveis dentre os próximos 90	Numérico	0,0%
Disponibilidade, requisitos e preços	28	Dias disponíveis dentre os próximos 365	Numérico	0,0%
Disponibilidade, requisitos e preços	29	Disponível para reserva imediata	Boleano	0,0%
Disponibilidade, requisitos e preços	30	Política de cancelamento	Categórico (6 categorias)	0,0%
Disponibilidade, requisitos e preços	31	Requisição de foto de quem está alugando	Boleano	0,0%
Disponibilidade, requisitos e preços	32	Requer telefone verificado de quem está alugando	Boleano	0,0%
Imóvel	33	Tipo de acomodação	Categórico (36 categorias)	0,0%
Imóvel	34	Tipo de quarto	Categórico (4 categorias)	0,0%
Imóvel	35	Número de hóspedes que o imóvel acomoda	Numérico	0,0%
Imóvel	36	Números de banheiros	Numérico	0,2%
Imóvel	37	Número de quartos	Numérico	0,2%
Imóvel	38	Número de camas	Numérico	0,2%

Imóvel	39	Tipo de cama	Categórico (5 categorias)	0,0%
Imóvel	40	Comodidades	Lista (172 categorias)	0,0%
Imóvel	41	Hóspedes inclusos	Numérico	0,0%
Localização	42	Nome do bairro tratado	Categórico (154 categorias)	0,4%
Localização	43	Latitude do imóvel	Numérico	0,0%
Localização	44	Longitude do imóvel	Numérico	0,0%
Localização	45	Localização fornecida é exata?	Boleano	0,0%
Localização	46	Distância em quilômetros até: Bondinho Pão de Açúcar	Numérico	0,0%
Localização	47	Distância em quilômetros até: Corcovado - Cristo Redentor	Numérico	0,0%
Localização	48	Distância em quilômetros até: Jardim Botânico	Numérico	0,0%
Localização	49	Distância em quilômetros até: Morro da Urca	Numérico	0,0%
Localização	50	Distância em quilômetros até: Praia de Ipanema	Numérico	0,0%
Localização	51	Distância em quilômetros até: Praia da Barra da Tijuca	Numérico	0,0%
Localização	52	Distância em quilômetros até: Maracanã	Numérico	0,0%
Localização	53	Distância em quilômetros até: Pedra Bonita	Numérico	0,0%
Localização	54	Distância em quilômetros até: Museu do Amanhã	Numérico	0,0%
Localização	55	Distância em quilômetros até: Praia do Arpoador	Numérico	0,0%
Localização	56	Distância em quilômetros até: Praia de Copacabana	Numérico	0,0%
Localização	57	Distância em quilômetros até: Parque Lage	Numérico	0,0%
Localização	58	Distância em quilômetros até: Prainha Beach	Numérico	0,0%
Localização	59	Distância em quilômetros até: Praia do Grumari	Numérico	0,0%

Localização	60	Distância em quilômetros até: AquaRio	Numérico	0,0%
Localização	61	Distância em quilômetros até: Parque Nacional da Tijuca	Numérico	0,0%
Localização	62	Distância em quilômetros até: Praia do Leblon	Numérico	0,0%
Localização	63	Distância em quilômetros até: Real Gabinete Português da Leitura	Numérico	0,0%
Localização	64	Distância em quilômetros até: Theatro Municipal do Rio de Janeiro	Numérico	0,0%
Localização	65	Distância em quilômetros até: Escadaria Selarón	Numérico	0,0%
Reputação	66	Número de comentários	Numérico	0,0%
Reputação	67	Número de comentários no último ano	Numérico	0,0%
Reputação	68	Nota geral	Numérico	44,0%
Reputação	69	Nota de precisão	Numérico	44,1%
Reputação	70	Nota da limpeza	Numérico	44,0%
Reputação	71	Nota do check-in	Numérico	44,1%
Reputação	72	Nota da comunicação	Numérico	44,0%
Reputação	73	Nota da localização	Numérico	44,0%
Reputação	74	Nota do valor	Numérico	44,0%
Reputação	75	Número de comentários por mês	Numérico	41,5%
Reputação	76	Percentual dos comentários em português	Numérico	41,5%
Reputação	77	Percentual dos comentários em inglês	Numérico	41,5%
Reputação	78	Percentual dos comentários em espanhol	Numérico	41,5%
Reputação	79	Percentual dos comentários em outros idiomas que não português, inglês e espanhol	Numérico	41,5%
