Teste de conhecimento Cinnecta

Claudio Resende
Vaga: cientista de dados
09/10/2020

Introdução

Este documento apresenta uma análise descritiva e preditiva de dados de acomodações do AirBnB. A base de dados fornecida contém 34 variáveis (colunas) e 7146 observações (linhas).

Para a análise aqui realizada foram selecionadas 22 colunas: foram excluídas colunas identificadoras da acomodação, como latitude e longitude, e as colunas *booleanas* relacionadas às avaliações (colunas '_na').

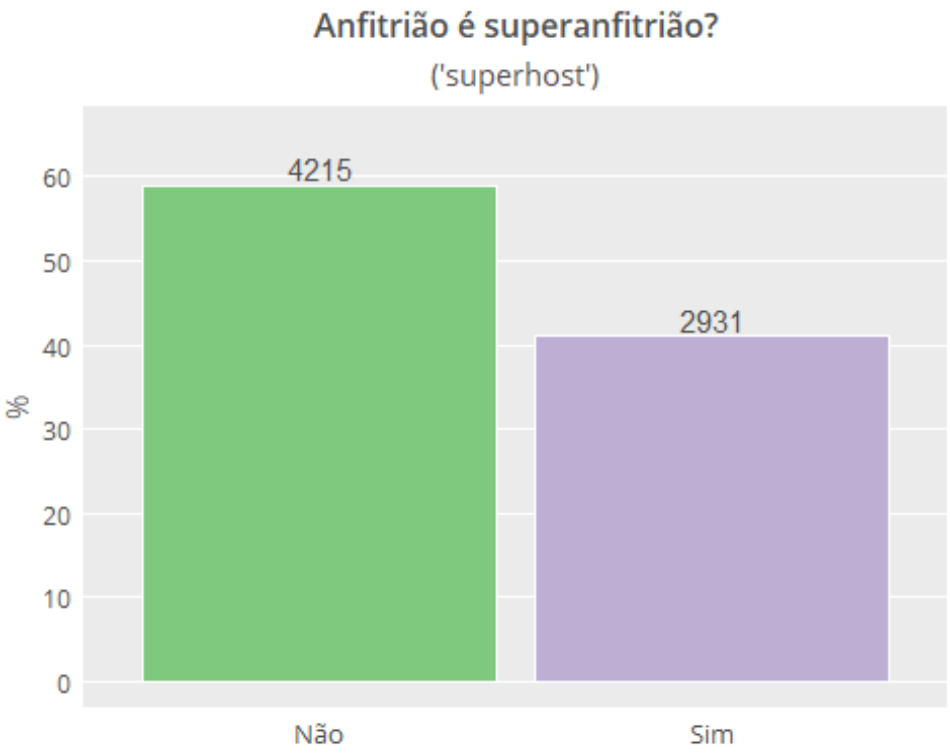
A primeira parte da análise consiste em explorar as variáveis para identificar eventuais padrões, tendências, vieses e outros tipos de comportamento dos dados que possam requerer transformações. Em seguida, são propostos modelos estatísticos para analisar a relação entre as variáveis.

Análise descritiva/exploratória

A seguir as variáveis da base de dados são analisadas individualmente e, em seguinda, em combinação com uma ou duas outras variáveis.

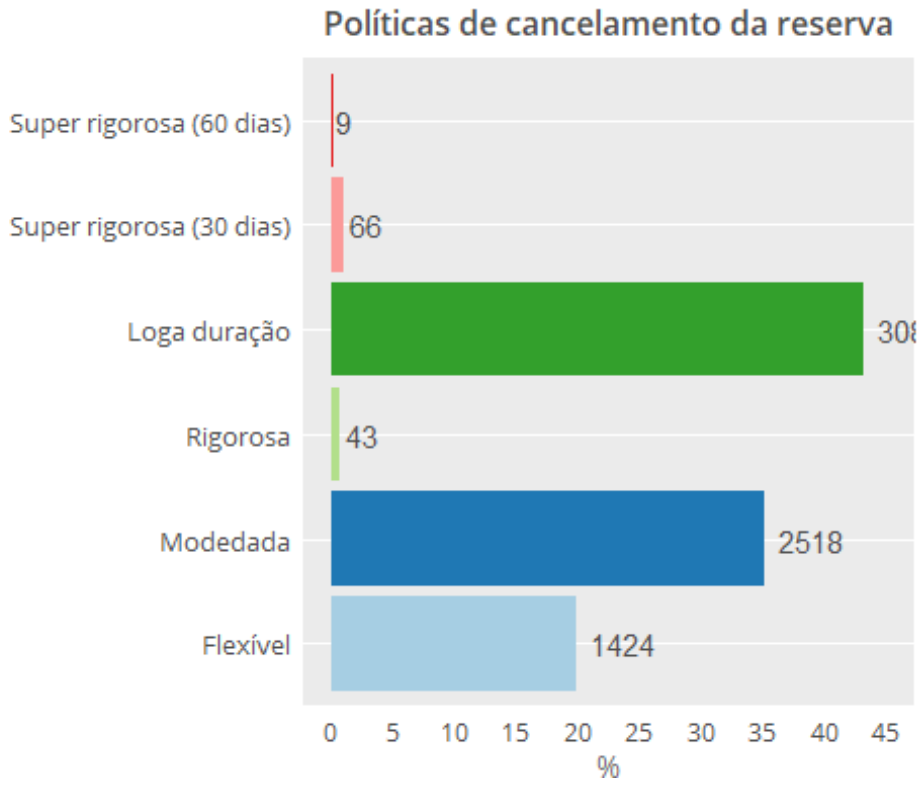
Anfitrião

O tipo de anfitrião é uma variável importante nos serviços oferecidos pelo AirBnB porque indicam o nível de experiência do anfitrião e a qualidade do serviço prestado. A base de dados analisada está distribuída em 59% e 41% para anfitriões comuns (host) e superanfitriões (superhost), respectivamente.



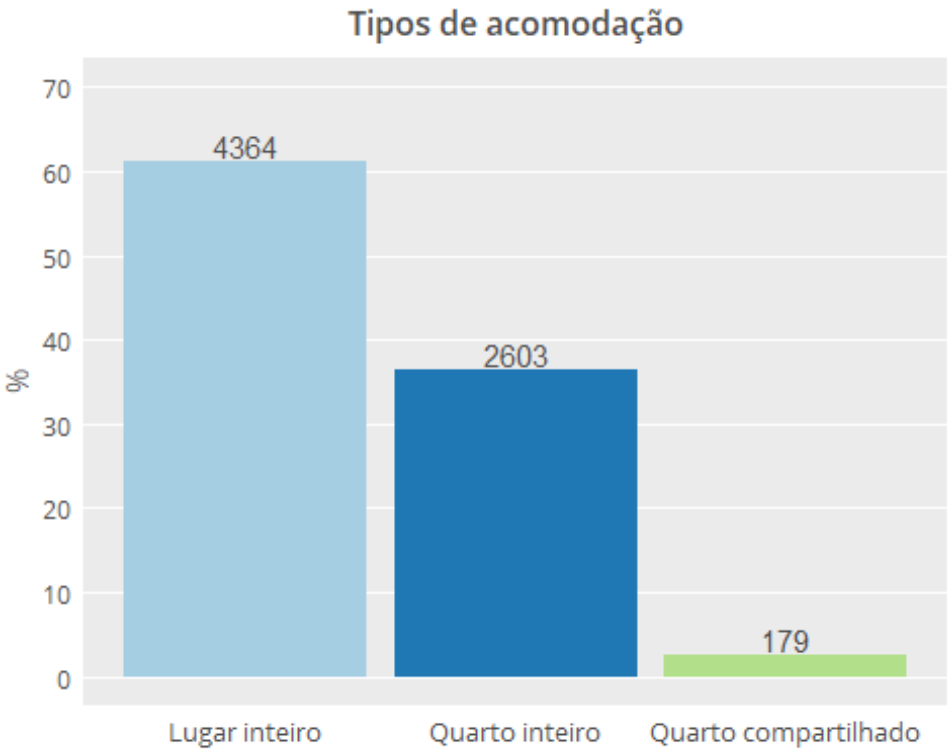
Políticas de cancelamento

O tipo de política de cancelamento possui uma distribuição mais concentrada. O AirBnB oferece seis tipos de política para o anfitrião escolher. Entre os dados disponinilizados, três categorias somam 98,3% dos casos: flexível, moderada e longa duração. Como poderá ser visto adiante, essas políticas se relacionam com o tipo de acomodação oferecida. Por exemplo, acomodações completas são propícias para estadias longas, ao passo que estadias simples, como quartos, são ideais para estadias curtas de poucos dias.

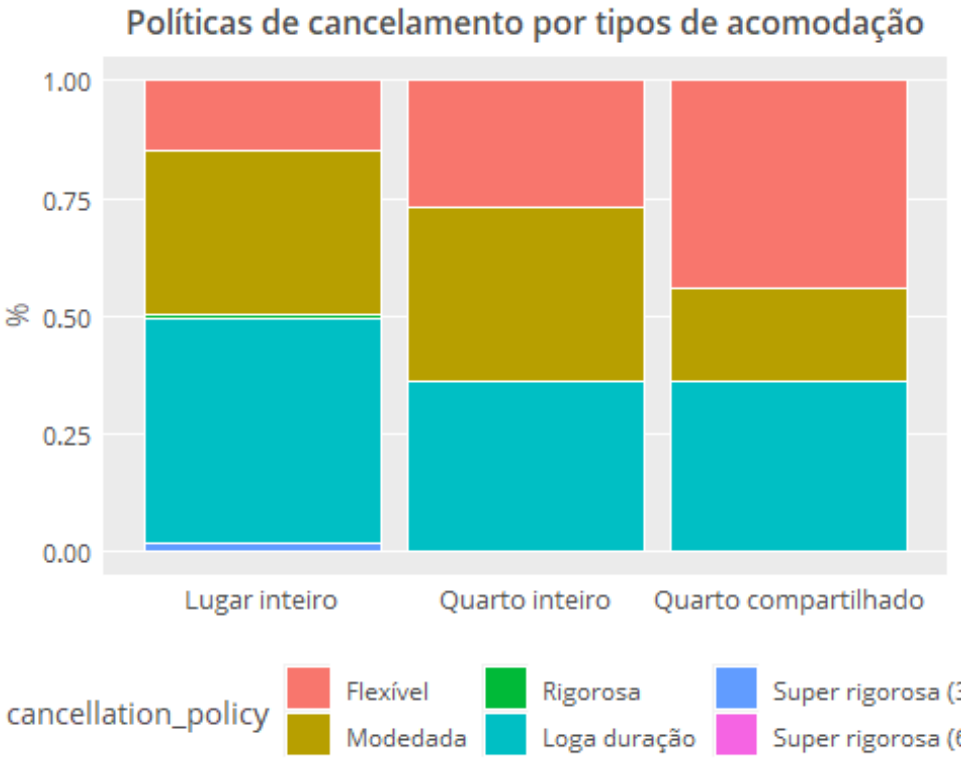


Tipos de acomodação

Lugares inteiros são a maioria das ofertas presentes na base de dados 61.1%, seguido de quartos inteiros 36.4%. Quartos compartilhados são apenas 2.5%.

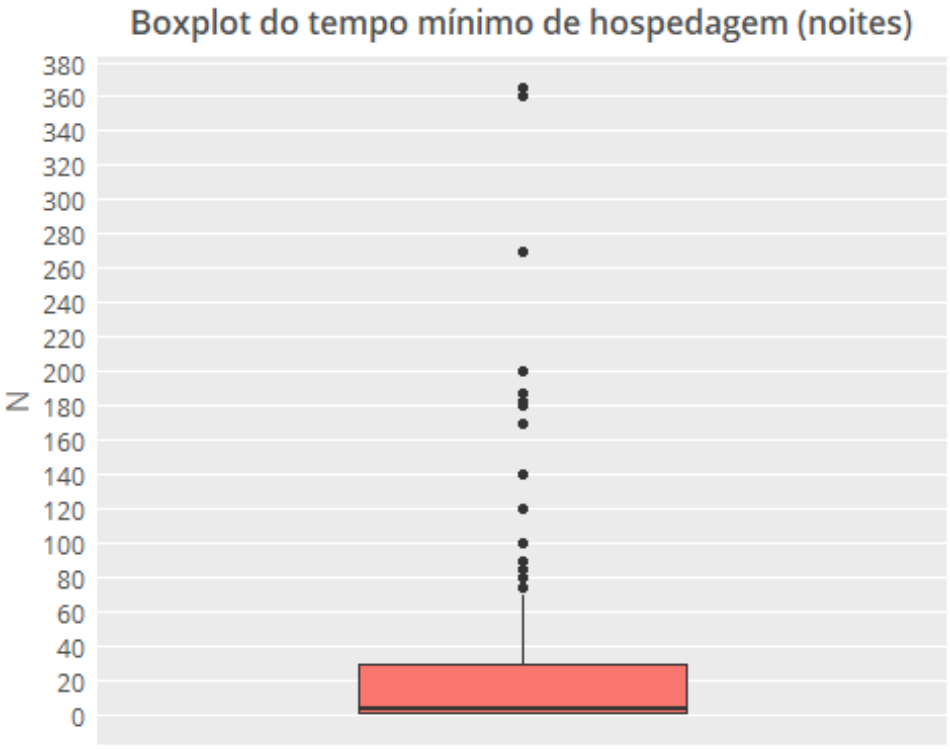


Quartos compartilhados, pela sua natureza, além de serem minoria na base, são também os mais permissivos quanto à política de cancelamento.



Mínimo de noites

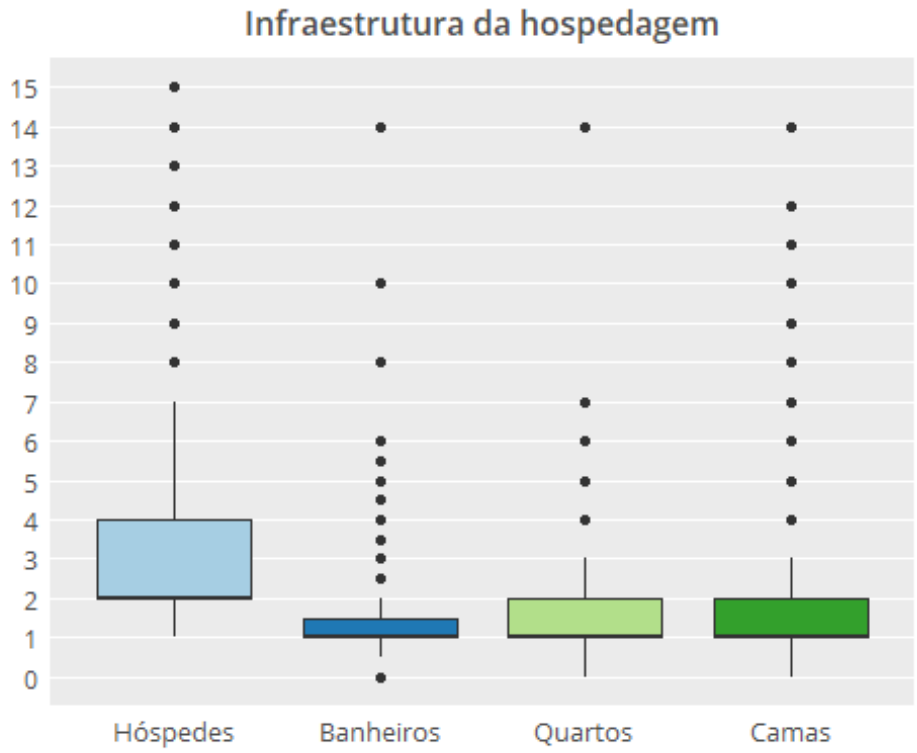
Outra variável intrinsecamente relacionada com o tipo de acomodação é o número mínimo de noites que o hóspede deve contratar o serviço. Como pode ser visto no boxplot abaixo, 75% das acomodações exigem até 30 dias de tempo mínimo, sendo que a mediana é de quatro dias.



Infraestrutura da hospedagem

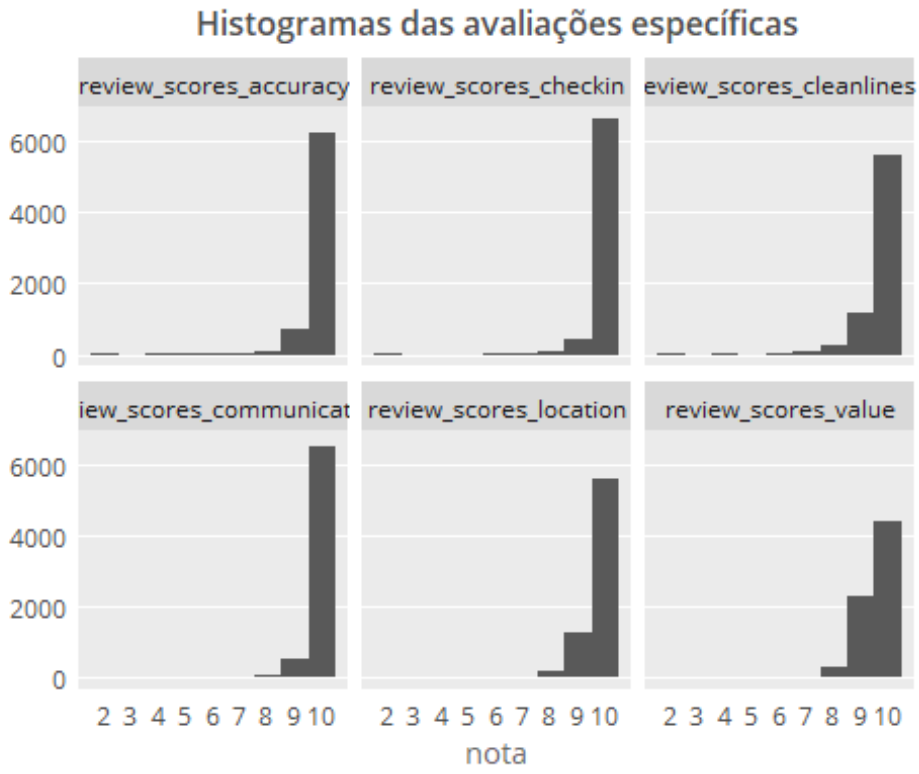
Os dados de infraestrutura também são importantes para avaliar as diferenças entre as acomodações oferecidas. Uma das variáveis disponíveis, tipo de cama (`dados$bed_type`), varia pouco (99% das camas são do tipo 'cama de verdade', ou 'real bed'), e portanto não será considerada na análise.

Por outro lado, os números de cômodos (banheiros e quartos) e de camas, bem como o número máximo de hóspedes, variam entre as hospedagens. O boxplot abaixo apresenta a distribuição de cada uma dessas variáveis: é possível perceber que há uma concentração em valores pequenos e a presença de *outliers*.

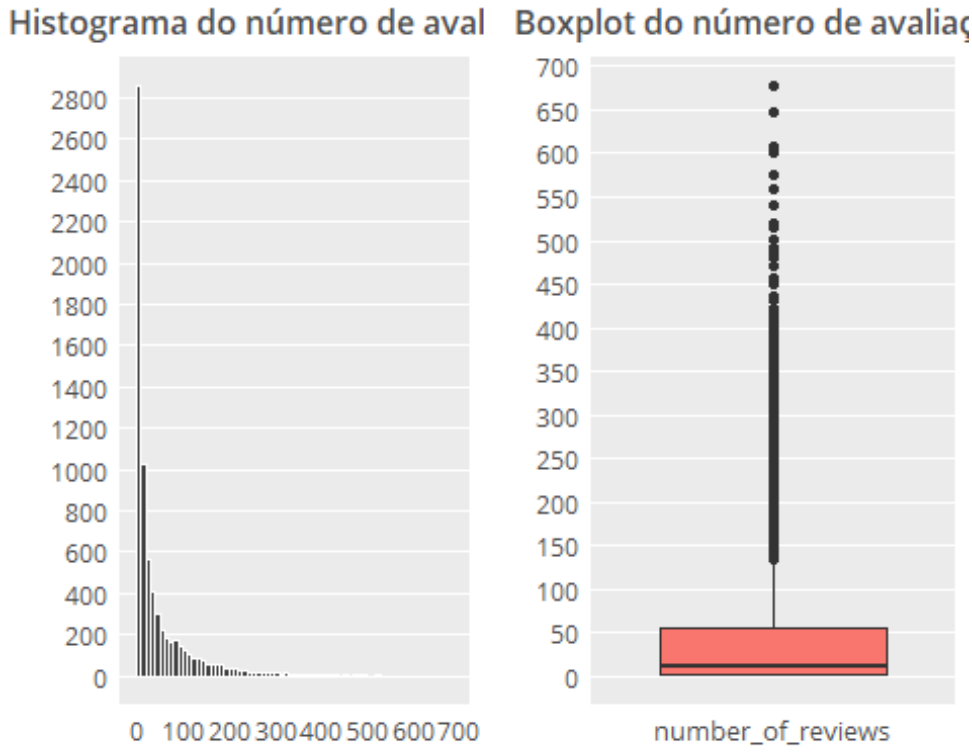


Avaliações

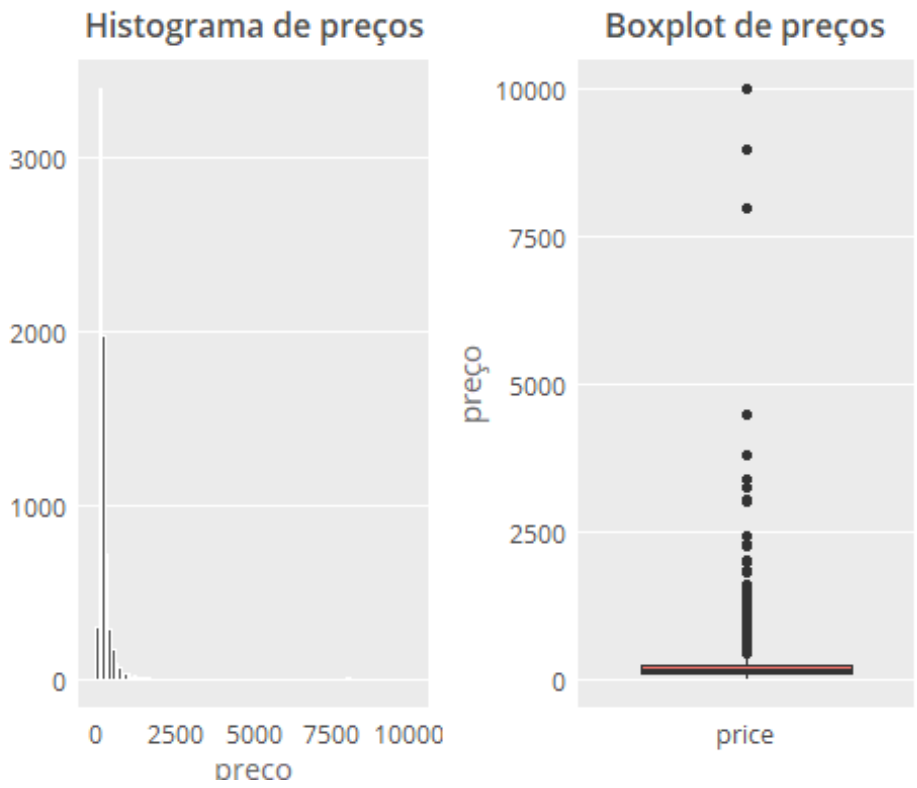
O número total de avaliações por hospedagem, o valor final da avaliação e valores das avaliações por item também possuem uma forte concentração: a maioria das hospedagens possui poucas ou nenhuma avaliação, e são também, em sua maioria, muito bem avaliadas (médias acima de 9,5), conforme pode ser visto nos histogramas abaixo.



A maioria das hospedagens possui poucas avaliações: a média é de 43.6 avaliações e a mediana é de 11. O histograma e o boxplot abaixo mostram essa concentração e permitem verificar também que há *outliers* na variável

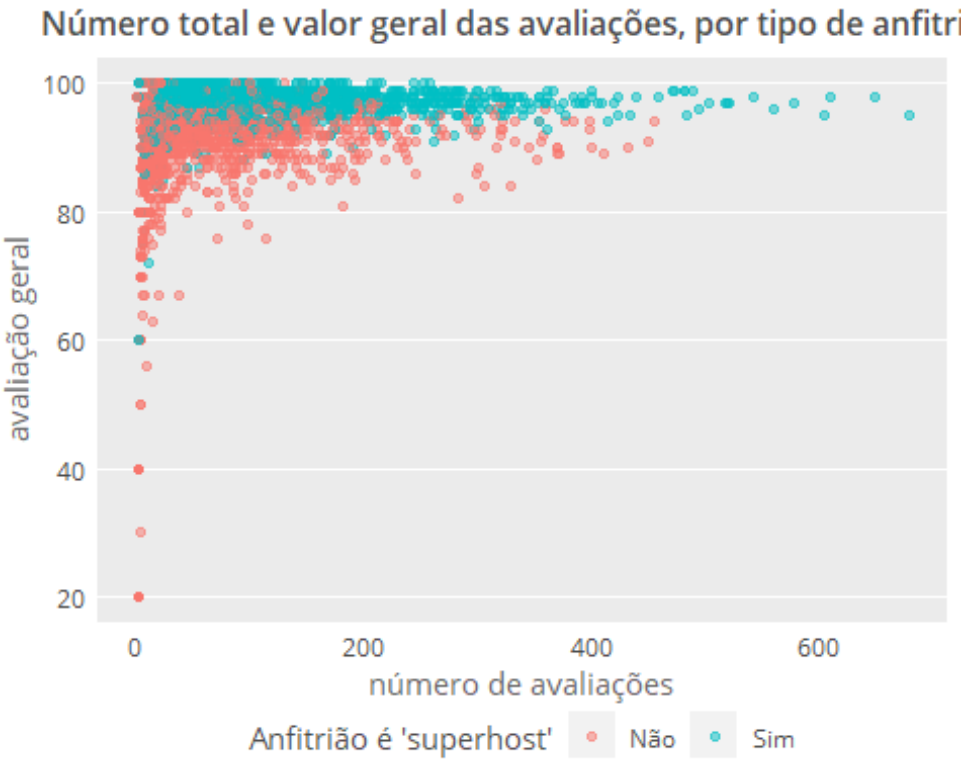


Comportamento semelhante aos dos preços praticados. O valor médio das hospedagens é U$213.3, mas há hospedagens que cobram U$10.000.

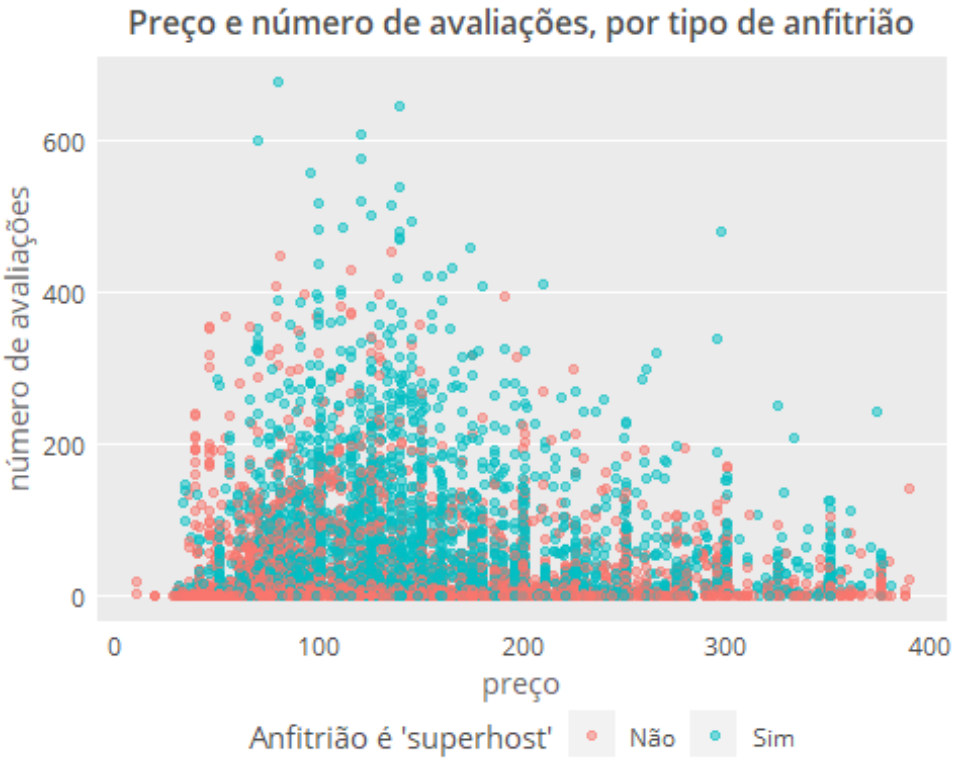


O que diferencia 'Superhosts' dos anfitriões comuns?

As análises a seguir sugerem que superhosts possuem mais e melhores avaliações do que anfitriões comuns.

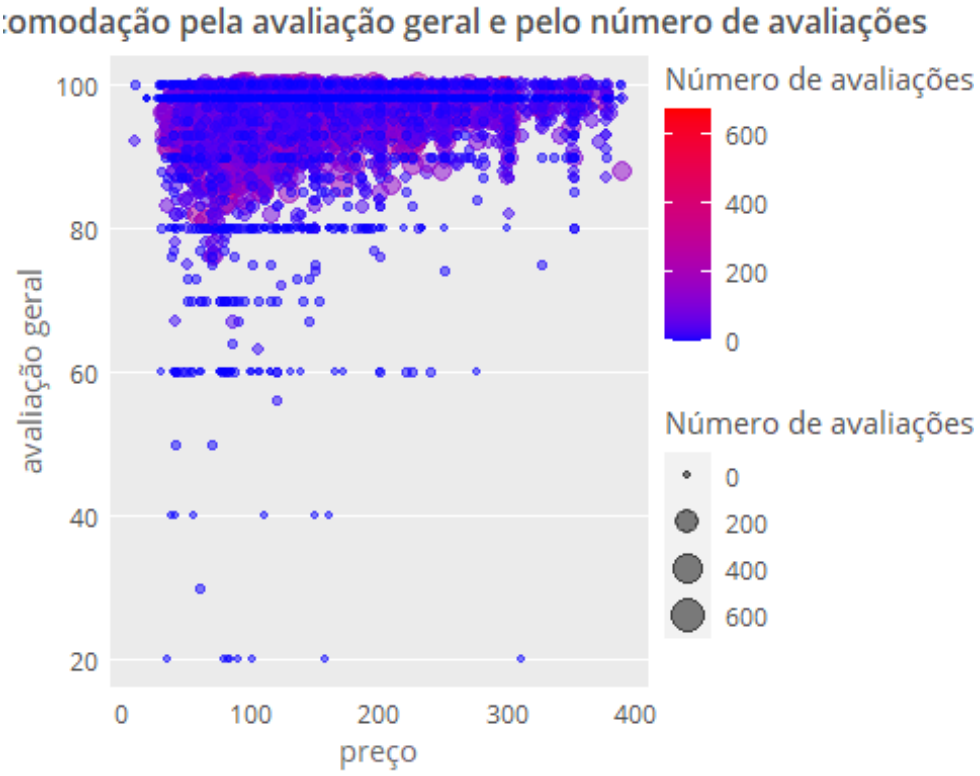


Como como pode ser visto a seguir, a maioria dos superhosts pratica preços abaixo da média (U$213.3).



Avaliações e preços

As hospedagens com maior número de avaliações são aquelas com menor preço. Ao mesmo tempo, as hospedagens com avaliações mais baixas são majoritariamente as mais baratas.



Predição

Sabendo como as variáveis se comportam, de maneira geral, podemos agora empreender uma análise preditiva. Proponho aqui dois modelos gerais: *Avaliação geral como variável dependente (resposta);* Preço como variável dependente (resposta).

O primeiro modelo geral foi gerado inicialmente com todas as variáveis como independentes (explicativas). O objetivo dessa etapa é o de identificar quais variáveis possuem significância na explicação das variações na avaliação geral. Nesse modelo, as variáveis 'price', 'host_is_superhost', 'property_type', 'cancellation_policy', 'room_type', 'bath_rooms', 'bed_rooms', 'beds', 'minimum_nights' e 'number_of_reviews' apresentaram significância.

Assim, gerou-se um novo modelo apenas com essas variáveis. O resultado está apresentado no quadro abaixo. O resultado geral do modelo, embora significativo, é

pouco satisfatório na sua capacidade explicativa. O valor de R2 ajustado (0,06119) indica uma capacidade explicativa do modelo pequena. Ou seja, as notas finais das hospedagens são explicadas por fatores outros, certamente aqueles que possuem avaliações específicas: limpeza, serviço de checkin, localização, comunicação etc.

```
summary(modelo2)
```

```
##
## Call:
## lm(formula = review_scores_rating ~ price + host_is_superhost +
##     room_type + bathrooms + bedrooms + beds + minimum_nights +
##     number_of_reviews, data = modelo)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -75.347  -1.016   1.435   2.970  11.130
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)           95.7160938  0.2159045 443.326  < 2e-16 ***
## price                  0.0010514  0.0002521   4.171 3.07e-05 ***
## host_is_superhost1     2.5336791  0.1546655  16.382  < 2e-16 ***
## room_typePrivate room -0.7540517  0.1663220  -4.534 5.89e-06 ***
## room_typeShared room  -0.5553824  0.4908922  -1.131 0.257936
## bathrooms             -0.3815666  0.1046278  -3.647 0.000267 ***
## bedrooms               0.8281040  0.1219552   6.790 1.21e-11 ***
## beds                  -0.3942144  0.0937858  -4.203 2.66e-05 ***
## minimum_nights        -0.0193460  0.0033923  -5.703 1.23e-08 ***
## number_of_reviews     -0.0059757  0.0010713  -5.578 2.52e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.091 on 7136 degrees of freedom
## Multiple R-squared:  0.06237,    Adjusted R-squared:  0.06119
## F-statistic: 52.74 on 9 and 7136 DF,  p-value: < 2.2e-16
```

O segundo modelo geral proposto foi construído para predizer o preço do imóvel a partir das demais variáveis. Como feito anteriormente, primeiro gerou-se um modelo com todas as variáveis para determinar quais possuem significância. A seguir, apresenta-se o resultado do modelo com as variáveis 'review_scores_rating', 'room_type', 'accommodates', 'bathrooms', 'bedrooms', 'minimum_nights' e 'number_of_reviews'.

O modelo indica que acomodações de tipo 'quarto privado' e 'quarto compartilhado' são mais baratas que acomodações de tipo 'lugar inteiro' (coeficiente negativo). O mesmo ocorre com o número mínimo de noites e com o número de avaliações.

Por outro lado, os preços tendem a subir quanto maiores forem as notas, o número máximo de hóspedes, de banheiros e de quartos. Naturalmente, locais maiores tendem a ser mais caros (o que é esperado), da mesma forma que os locais mais bem avaliados tendem a ser mais procurados, o que contribui para elevar os preços.

Esse modelo possui uma capacidade explicativa razoável. O R2 ajustado de 0,1677 indica que o modelo explica 16,7% das variações de preços.

```
summary(modelo4)
```

```
##
## Call:
## lm(formula = price ~ review_scores_rating + room_type + accommodates +
##     bathrooms + bedrooms + minimum_nights + number_of_reviews,
##     data = modelo)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1369.8   -70.6   -23.7    26.5  9636.1
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -165.61599   53.34521  -3.105 0.001913 **
## review_scores_rating     2.32465    0.54080   4.299 1.74e-05 ***
## room_typePrivate room  -51.71239    8.04317  -6.429 1.36e-10 ***
## room_typeShared room  -129.38596   22.82057  -5.670 1.49e-08 ***
## accommodates            30.42277    2.86360  10.624  < 2e-16 ***
## bathrooms               22.33345    4.87963   4.577 4.80e-06 ***
## bedrooms                52.74687    5.73288   9.201  < 2e-16 ***
## minimum_nights          -0.59114    0.15835  -3.733 0.000191 ***
## number_of_reviews       -0.24710    0.04817  -5.130 2.97e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 284.1 on 7137 degrees of freedom
## Multiple R-squared:  0.1687, Adjusted R-squared:  0.1677
## F-statistic:   181 on 8 and 7137 DF,  p-value: < 2.2e-16
```

Os resultados, no entanto, não foram muito satisfatórios. Calculando os valores preditos e comparando-os com os valores originais, percebe-se que há diferenças significativas entre os valores, indicando que o modelo utilizado não é capaz de explicar adequadamente as variações nos preços. O MSE (erro quadrático médio) do modelo é de 283, um valor muito alto.

```
modelo %>% select(price, predito_modelo4) %>% head(15)
```

```
##     price predito_modelo4
##  1:   170       181.15364
##  2:   235       296.97783
```

```
##  3:    65       160.07569
##  4:    65       180.89680
##  5:   785       340.17282
##  6:   255       345.12361
##  7:   139        16.36843
##  8:   135        24.58520
##  9:   265       287.06736
## 10:   177       286.24301
## 11:   194       379.75468
## 12:   139       172.91504
## 13:    85       151.17059
## 14:    85        91.88988
## 15:    79        17.84628
```

Conclusões

Os modelos propostos não foram capazes de explicar as variações das avaliações gerais e dos preços das hospedagens. Seria preciso pensar outros modelos, com diferentes combinações de variáveis, para investigar se é possível predizer os valores de avaliação e preço a partir das variáveis presentes na base ou se seria necessário obter outras variáveis para isso.