

# Lista 2: Preço de acomodações no site *Airbnb*.

*Modelos lineares Generalizados*

*2/2019*

Nesta lista de exercícios utilizaremos modelos lineares generalizados para descrever o preço anunciado no site Airbnb, na cidade de Seattle/EUA, de diversas acomodações, considerando-se várias das principais características do imóvel/quarto. Os dados estão disponíveis em nossa pasta do Moodle e pelo link <https://www.kaggle.com/airbnb/seattle>.

O código a seguir reparte o conjunto de dados em *treinamento* e *teste* para viabilizar comparações entre os diferentes grupos. O objeto *covariaveis* lista o conjunto de informações de devem ser consideradas (mas não necessariamente incluídas no modelo final, obviamente).

```
### Carregando os pacotes exigidos
library(readr)
require(tidyverse)
set.seed(22019)

### Carregando os dados
listings <- read_csv("listings.csv")
calendar <- read_csv("calendar.csv")

### Limpando a base
calendar <- calendar %>%
  filter(!is.na(price)) %>% # filtrando os imóveis disponíveis
  mutate(price=str_replace($.price, "\\$", "")) %>%
  str_replace("\\\\", "\\") %>%
  as.numeric() # transformando os dados de preços em numérico

### Dividindo o banco de dados em treino e teste
n.obs <- nrow(calendar)
treino <- sample(n.obs, n.obs*.8)
teste <- setdiff(1:n.obs, treino)
calendar.treino <- calendar[treino, ]
calendar.teste <- calendar[teste, ]

### Definindo as informações a serem consideradas no modelo
covariaveis <- c("date", "host_since", "host_location", "host_response_time",
"host_response_rate", "host_acceptance_rate", "host_is_superhost",
"host_total_listings_count", "host_has_profile_pic", "host_identity_verified",
"neighbourhood_group_cleansed", "zipcode", "latitude", "longitude", "is_location_exact",
"property_type", "room_type", "accommodates", "bathrooms", "bedrooms", "beds",
"bed_type", "amenities", "square_feet", "security_deposit", "cleaning_fee",
"guests_included", "extra_people", "minimum_nights", "maximum_nights", "number_of_reviews",
"first_review", "last_review", "review_scores_rating", "review_scores_accuracy",
"review_scores_cleanliness", "review_scores_checkin", "review_scores_communication",
"review_scores_location", "review_scores_value", "requires_license", "instant_bookable",
"cancellation_policy", "require_guest_profile_picture", "require_guest_phone_verification",
"calculated_host_listings_count", "reviews_per_month")
```

**Crie um relatório que responda, da melhor maneira possível, os itens a seguir.**

- a) Ajuste um MLG para estimar a probabilidade de um dado apartamento (isso é, com um nível fixo das covariáveis) ser anunciado por um valor superior a \$200. Qual é o erro quadrático médio do modelo no conjunto de teste?
- b) Considerando fixas as demais covariáveis, qual é o ponto ou região da cidade com maiores valores anunciados?
- c) Descreva a influência da data no preço dos alugueis.
- d) Quais são, em ordem decrescente, os cinco principais fatores na determinação do preço das unidades ofertadas?
- e) De que modo o preço é impactado pelo aumento das notas dos *reviews*?
- f) Enumere as suposições feitas pelo modelo, comente a razoabilidade de cada uma delas e indique formas de melhorar o modelo (não é preciso implementar as estratégias elencadas).

**Atenção:** Formate o texto com esmero. Uma boa apresentação é fundamental para a qualidade do trabalho.