Airbnb - Revenue Management

David Gómez Sedas

A continuación se enumeran todas las librerías utilizadas en el AB:

library(tidyverse)

── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
✔ dplyr 1.1.4 ✔ readr 2.1.5  
✔ forcats 1.0.0 ✔ stringr 1.5.1  
✔ ggplot2 3.4.4 ✔ tibble 3.2.1  
✔ lubridate 1.9.3 ✔ tidyr 1.3.0  
✔ purrr 1.0.2   
── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
✖ dplyr::filter() masks stats::filter()  
✖ dplyr::lag() masks stats::lag()  
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(tidymodels)

── Attaching packages ────────────────────────────────────── tidymodels 1.1.1 ──  
✔ broom 1.0.5 ✔ rsample 1.2.0  
✔ dials 1.2.0 ✔ tune 1.1.2  
✔ infer 1.0.5 ✔ workflows 1.1.3  
✔ modeldata 1.3.0 ✔ workflowsets 1.0.1  
✔ parsnip 1.1.1 ✔ yardstick 1.3.0  
✔ recipes 1.0.9   
── Conflicts ───────────────────────────────────────── tidymodels\_conflicts() ──  
✖ scales::discard() masks purrr::discard()  
✖ dplyr::filter() masks stats::filter()  
✖ recipes::fixed() masks stringr::fixed()  
✖ dplyr::lag() masks stats::lag()  
✖ yardstick::spec() masks readr::spec()  
✖ recipes::step() masks stats::step()  
• Use suppressPackageStartupMessages() to eliminate package startup messages

library(quarto)  
library(fastDummies)

Thank you for using fastDummies!  
To acknowledge our work, please cite the package:  
Kaplan, J. & Schlegel, B. (2023). fastDummies: Fast Creation of Dummy (Binary) Columns and Rows from Categorical Variables. Version 1.7.1. URL: https://github.com/jacobkap/fastDummies, https://jacobkap.github.io/fastDummies/.

library(GGally)

Registered S3 method overwritten by 'GGally':  
 method from   
 +.gg ggplot2

library(factoextra)

Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa

# Datos usados

df <- read\_csv("listings\_bcn.csv", show\_col\_types = FALSE)

## Origen de los datos

Los datos han sido extraidos de “Inside Airbnb”, el cual es un proyecto que provee datos e información sobre el impacto que tiene Airbnb sobre las comunidades residenciales.

La finalidad de este proyecto es trabajar en torno a una visión que consiste en informar a estas comunidades sobre el uso real de airbnb.

Los datos son extraidos de la web de Airbnb oficial y son almacenados en archivos compatibles con datos tabulares para distribuirlos. Podemos encontrar que los datos están separados por ciudades y estos se suelen actualizar frecuentemente para mostrar una imagen lo mas actual posible.

Todos los datos de la web estan licenciados bajo Creative Commons Attribution 4.0 International License.

Concretamente, los datos utilizados en este trabajo consisten en todas las residencias activas en Airbnb de la ciudad de Barcelona. En ellas podemos encontar diferentes tipos de servicios: Apartamentos, casas, habitaciones y camas en habitaciones compartidas.

## Contenido

Los datos utilizados consisten en unos datos tabulares los cuales consisten en 18086 entradas y 75 columnas.

str(df)

spc\_tbl\_ [18,086 × 75] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
 $ id : num [1:18086] 18674 23197 117010 32711 118228 ...  
 $ listing\_url : chr [1:18086] "https://www.airbnb.com/rooms/18674" "https://www.airbnb.com/rooms/23197" "https://www.airbnb.com/rooms/117010" "https://www.airbnb.com/rooms/32711" ...  
 $ scrape\_id : num [1:18086] 2.02e+13 2.02e+13 2.02e+13 2.02e+13 2.02e+13 ...  
 $ last\_scraped : Date[1:18086], format: "2023-09-06" "2023-09-06" ...  
 $ source : chr [1:18086] "city scrape" "city scrape" "previous scrape" "city scrape" ...  
 $ name : chr [1:18086] "Rental unit in Barcelona · ★4.30 · 3 bedrooms · 6 beds · 2 baths" "Rental unit in Sant Adria de Besos · ★4.77 · 3 bedrooms · 4 beds · 2 baths" "Rental unit in Barcelona · ★4.55 · 3 bedrooms · 6 beds · 2 baths" "Rental unit in Barcelona · ★4.46 · 2 bedrooms · 3 beds · 1.5 baths" ...  
 $ description : chr [1:18086] "110m2 apartment to rent in Barcelona. Located in the Eixample district, near the Sagrada Familia. It has a smal"| \_\_truncated\_\_ "Beautiful spacious apartment, large terrace, 5 minutes walk from the CCIB center, sea, Port Forum.<br />Excelle"| \_\_truncated\_\_ "Have an authentic Barcelona experience! Modern 100m2 recently decorated apartment - definitely one to consider "| \_\_truncated\_\_ "A lovely two bedroom apartment only 250 m from Barcelona's iconic Sagrada Familia.<br /><br /><b>The space</b><"| \_\_truncated\_\_ ...  
 $ neighborhood\_overview : chr [1:18086] "Apartment in Barcelona located in the heart of Eixample district, within only 150 m form the great Sagrada Fami"| \_\_truncated\_\_ "Strategically located in the Parc del Fòrum, a spacious area where all kinds of music festivals and events are "| \_\_truncated\_\_ "Apartment is located just 240 meters from Sagrada Familia. Right on Avenida Gaudí, a cozy boulevard that leads "| \_\_truncated\_\_ "What's nearby <br />This apartment is located in central Barcelona only 250m from the famous Sagrada Familia. "| \_\_truncated\_\_ ...  
 $ picture\_url : chr [1:18086] "https://a0.muscache.com/pictures/13031453/413cdbfc\_original.jpg" "https://a0.muscache.com/pictures/miso/Hosting-23197/original/0820d261-5f12-4ac9-8604-fefa1c5ec62a.jpeg" "https://a0.muscache.com/pictures/4052d8b5-7352-41be-821d-4188eb99a89a.jpg" "https://a0.muscache.com/pictures/357b25e4-f414-48c5-aad0-e89342b1daa2.jpg" ...  
 $ host\_id : num [1:18086] 71615 90417 567180 135703 567180 ...  
 $ host\_url : chr [1:18086] "https://www.airbnb.com/users/show/71615" "https://www.airbnb.com/users/show/90417" "https://www.airbnb.com/users/show/567180" "https://www.airbnb.com/users/show/135703" ...  
 $ host\_name : chr [1:18086] "Mireia And Maria" "Etain (Marnie)" "Sandra" "Nick" ...  
 $ host\_since : Date[1:18086], format: "2010-01-19" "2010-03-09" ...  
 $ host\_location : chr [1:18086] "Barcelona, Spain" "Catalonia, Spain" "Barcelona, Spain" "Barcelona, Spain" ...  
 $ host\_about : chr [1:18086] "We are Mireia (47) & Maria (49), two multilingual entrepreneurs loving Barcelona and having big experience in t"| \_\_truncated\_\_ "Hi there, I'm marnie from Australia, though I have been living in Spain for many years.\nMy lovely co-host, Jak"| \_\_truncated\_\_ "Suite Home Sagrada Familia is a complex of apartments with 12 years of experience. \"Welcome to your home” is o"| \_\_truncated\_\_ "I'm Nick your English host in Barcelona.\r\n\r\nI'm a full time host with several holiday rental apartments tha"| \_\_truncated\_\_ ...  
 $ host\_response\_time : chr [1:18086] "within an hour" "within an hour" "within a few hours" "within an hour" ...  
 $ host\_response\_rate : chr [1:18086] "97%" "100%" "88%" "100%" ...  
 $ host\_acceptance\_rate : chr [1:18086] "90%" "94%" "98%" "100%" ...  
 $ host\_is\_superhost : logi [1:18086] FALSE TRUE FALSE FALSE FALSE FALSE ...  
 $ host\_thumbnail\_url : chr [1:18086] "https://a0.muscache.com/im/users/71615/profile\_pic/1426612511/original.jpg?aki\_policy=profile\_small" "https://a0.muscache.com/im/pictures/user/44b56b17-e3c1-4fc9-a036-09edc5a0bbee.jpg?aki\_policy=profile\_small" "https://a0.muscache.com/im/pictures/user/7aaeb41d-1ee2-48cb-927a-c39c0868eb7d.jpg?aki\_policy=profile\_small" "https://a0.muscache.com/im/users/135703/profile\_pic/1346864531/original.jpg?aki\_policy=profile\_small" ...  
 $ host\_picture\_url : chr [1:18086] "https://a0.muscache.com/im/users/71615/profile\_pic/1426612511/original.jpg?aki\_policy=profile\_x\_medium" "https://a0.muscache.com/im/pictures/user/44b56b17-e3c1-4fc9-a036-09edc5a0bbee.jpg?aki\_policy=profile\_x\_medium" "https://a0.muscache.com/im/pictures/user/7aaeb41d-1ee2-48cb-927a-c39c0868eb7d.jpg?aki\_policy=profile\_x\_medium" "https://a0.muscache.com/im/users/135703/profile\_pic/1346864531/original.jpg?aki\_policy=profile\_x\_medium" ...  
 $ host\_neighbourhood : chr [1:18086] "la Sagrada Família" "El Besòs i el Maresme" "la Sagrada Familia" "Camp d'en Grassot i Gràcia Nova" ...  
 $ host\_listings\_count : num [1:18086] 47 6 19 3 19 1 34 3 34 2 ...  
 $ host\_total\_listings\_count : num [1:18086] 48 9 19 15 19 1 50 4 50 2 ...  
 $ host\_verifications : chr [1:18086] "['email', 'phone']" "['email', 'phone']" "['email', 'phone']" "['email', 'phone', 'work\_email']" ...  
 $ host\_has\_profile\_pic : logi [1:18086] TRUE TRUE TRUE TRUE TRUE TRUE ...  
 $ host\_identity\_verified : logi [1:18086] TRUE TRUE FALSE TRUE FALSE TRUE ...  
 $ neighbourhood : chr [1:18086] "Barcelona, CT, Spain" "Sant Adria de Besos, Barcelona, Spain" "Barcelona, Catalonia, Spain" "Barcelona, Catalonia, Spain" ...  
 $ neighbourhood\_cleansed : chr [1:18086] "la Sagrada Família" "el Besòs i el Maresme" "la Sagrada Família" "el Camp d'en Grassot i Gràcia Nova" ...  
 $ neighbourhood\_group\_cleansed : chr [1:18086] "Eixample" "Sant Martí" "Eixample" "Gràcia" ...  
 $ latitude : num [1:18086] 41.4 41.4 41.4 41.4 41.4 ...  
 $ longitude : num [1:18086] 2.17 2.22 2.17 2.17 2.17 ...  
 $ property\_type : chr [1:18086] "Entire rental unit" "Entire rental unit" "Entire rental unit" "Entire rental unit" ...  
 $ room\_type : chr [1:18086] "Entire home/apt" "Entire home/apt" "Entire home/apt" "Entire home/apt" ...  
 $ accommodates : num [1:18086] 8 5 8 6 8 2 7 9 7 5 ...  
 $ bathrooms : logi [1:18086] NA NA NA NA NA NA ...  
 $ bathrooms\_text : chr [1:18086] "2 baths" "2 baths" "2 baths" "1.5 baths" ...  
 $ bedrooms : num [1:18086] 3 3 3 2 3 1 3 4 3 2 ...  
 $ beds : num [1:18086] 6 4 6 3 5 1 5 6 4 3 ...  
 $ amenities : chr [1:18086] "[\"Refrigerator\", \"30\\\" TV\", \"Iron\", \"Hangers\", \"City skyline view\", \"Free washer \\u2013 In unit\""| \_\_truncated\_\_ "[\"Refrigerator\", \"Toaster\", \"Cleaning products\", \"Room-darkening shades\", \"Fire extinguisher\", \"Publ"| \_\_truncated\_\_ "[\"Refrigerator\", \"Clothing storage: closet and dresser\", \"Toaster\", \"Room-darkening shades\", \"Fire ext"| \_\_truncated\_\_ "[\"Refrigerator\", \"Fire extinguisher\", \"First aid kit\", \"Iron\", \"Hangers\", \"Paid parking off premises"| \_\_truncated\_\_ ...  
 $ price : chr [1:18086] "$202.00" "$255.00" "$331.00" "$171.00" ...  
 $ minimum\_nights : num [1:18086] 1 3 2 21 2 3 3 5 3 90 ...  
 $ maximum\_nights : num [1:18086] 1125 300 30 31 28 ...  
 $ minimum\_minimum\_nights : num [1:18086] 1 3 2 1 2 3 2 2 3 90 ...  
 $ maximum\_minimum\_nights : num [1:18086] 4 5 3 1 3 3 6 5 6 90 ...  
 $ minimum\_maximum\_nights : num [1:18086] 1125 1125 30 31 28 ...  
 $ maximum\_maximum\_nights : num [1:18086] 1125 1125 32 31 32 ...  
 $ minimum\_nights\_avg\_ntm : num [1:18086] 2.2 3.2 2 1 2.1 3 3.3 4.9 3.3 90 ...  
 $ maximum\_nights\_avg\_ntm : num [1:18086] 1125 1125 30.3 31 28.7 ...  
 $ calendar\_updated : logi [1:18086] NA NA NA NA NA NA ...  
 $ has\_availability : logi [1:18086] TRUE TRUE FALSE TRUE FALSE TRUE ...  
 $ availability\_30 : num [1:18086] 4 16 0 6 0 2 7 5 6 30 ...  
 $ availability\_60 : num [1:18086] 11 31 0 17 0 9 18 20 12 60 ...  
 $ availability\_90 : num [1:18086] 21 61 0 43 0 32 18 38 31 90 ...  
 $ availability\_365 : num [1:18086] 34 150 0 310 0 303 266 194 279 365 ...  
 $ calendar\_last\_scraped : Date[1:18086], format: "2023-09-06" "2023-09-06" ...  
 $ number\_of\_reviews : num [1:18086] 38 73 48 95 50 395 32 211 85 131 ...  
 $ number\_of\_reviews\_ltm : num [1:18086] 8 11 6 21 18 62 14 31 33 0 ...  
 $ number\_of\_reviews\_l30d : num [1:18086] 0 1 1 1 0 6 1 2 1 0 ...  
 $ first\_review : Date[1:18086], format: "2013-05-27" "2011-03-15" ...  
 $ last\_review : Date[1:18086], format: "2023-06-26" "2023-08-15" ...  
 $ review\_scores\_rating : num [1:18086] 4.3 4.77 4.55 4.46 4.56 4.86 4.56 4.53 4.43 4.64 ...  
 $ review\_scores\_accuracy : num [1:18086] 4.41 4.93 4.59 4.44 4.5 4.9 4.81 4.61 4.64 4.74 ...  
 $ review\_scores\_cleanliness : num [1:18086] 4.62 4.89 4.57 4.47 4.62 4.93 4.75 4.62 4.48 4.68 ...  
 $ review\_scores\_checkin : num [1:18086] 4.76 4.94 4.82 4.86 4.56 4.93 4.84 4.68 4.79 4.69 ...  
 $ review\_scores\_communication : num [1:18086] 4.65 4.99 4.91 4.85 4.74 4.9 4.78 4.69 4.81 4.8 ...  
 $ review\_scores\_location : num [1:18086] 4.78 4.6 4.86 4.86 4.9 4.72 4.81 4.69 4.88 4.83 ...  
 $ review\_scores\_value : num [1:18086] 4.27 4.64 4.59 4.52 4.52 4.82 4.38 4.43 4.32 4.56 ...  
 $ license : chr [1:18086] "HUTB-002062" "HUTB005057" "HUTB000331" "HUTB-001722" ...  
 $ instant\_bookable : logi [1:18086] TRUE FALSE FALSE TRUE FALSE FALSE ...  
 $ calculated\_host\_listings\_count : num [1:18086] 30 2 19 3 19 1 32 2 32 2 ...  
 $ calculated\_host\_listings\_count\_entire\_homes : num [1:18086] 30 2 19 3 19 1 32 2 32 2 ...  
 $ calculated\_host\_listings\_count\_private\_rooms: num [1:18086] 0 0 0 0 0 0 0 0 0 0 ...  
 $ calculated\_host\_listings\_count\_shared\_rooms : num [1:18086] 0 0 0 0 0 0 0 0 0 0 ...  
 $ reviews\_per\_month : num [1:18086] 0.3 0.48 0.33 0.64 0.34 2.7 0.28 1.34 0.67 0.9 ...  
 - attr(\*, "spec")=  
 .. cols(  
 .. id = col\_double(),  
 .. listing\_url = col\_character(),  
 .. scrape\_id = col\_double(),  
 .. last\_scraped = col\_date(format = ""),  
 .. source = col\_character(),  
 .. name = col\_character(),  
 .. description = col\_character(),  
 .. neighborhood\_overview = col\_character(),  
 .. picture\_url = col\_character(),  
 .. host\_id = col\_double(),  
 .. host\_url = col\_character(),  
 .. host\_name = col\_character(),  
 .. host\_since = col\_date(format = ""),  
 .. host\_location = col\_character(),  
 .. host\_about = col\_character(),  
 .. host\_response\_time = col\_character(),  
 .. host\_response\_rate = col\_character(),  
 .. host\_acceptance\_rate = col\_character(),  
 .. host\_is\_superhost = col\_logical(),  
 .. host\_thumbnail\_url = col\_character(),  
 .. host\_picture\_url = col\_character(),  
 .. host\_neighbourhood = col\_character(),  
 .. host\_listings\_count = col\_double(),  
 .. host\_total\_listings\_count = col\_double(),  
 .. host\_verifications = col\_character(),  
 .. host\_has\_profile\_pic = col\_logical(),  
 .. host\_identity\_verified = col\_logical(),  
 .. neighbourhood = col\_character(),  
 .. neighbourhood\_cleansed = col\_character(),  
 .. neighbourhood\_group\_cleansed = col\_character(),  
 .. latitude = col\_double(),  
 .. longitude = col\_double(),  
 .. property\_type = col\_character(),  
 .. room\_type = col\_character(),  
 .. accommodates = col\_double(),  
 .. bathrooms = col\_logical(),  
 .. bathrooms\_text = col\_character(),  
 .. bedrooms = col\_double(),  
 .. beds = col\_double(),  
 .. amenities = col\_character(),  
 .. price = col\_character(),  
 .. minimum\_nights = col\_double(),  
 .. maximum\_nights = col\_double(),  
 .. minimum\_minimum\_nights = col\_double(),  
 .. maximum\_minimum\_nights = col\_double(),  
 .. minimum\_maximum\_nights = col\_double(),  
 .. maximum\_maximum\_nights = col\_double(),  
 .. minimum\_nights\_avg\_ntm = col\_double(),  
 .. maximum\_nights\_avg\_ntm = col\_double(),  
 .. calendar\_updated = col\_logical(),  
 .. has\_availability = col\_logical(),  
 .. availability\_30 = col\_double(),  
 .. availability\_60 = col\_double(),  
 .. availability\_90 = col\_double(),  
 .. availability\_365 = col\_double(),  
 .. calendar\_last\_scraped = col\_date(format = ""),  
 .. number\_of\_reviews = col\_double(),  
 .. number\_of\_reviews\_ltm = col\_double(),  
 .. number\_of\_reviews\_l30d = col\_double(),  
 .. first\_review = col\_date(format = ""),  
 .. last\_review = col\_date(format = ""),  
 .. review\_scores\_rating = col\_double(),  
 .. review\_scores\_accuracy = col\_double(),  
 .. review\_scores\_cleanliness = col\_double(),  
 .. review\_scores\_checkin = col\_double(),  
 .. review\_scores\_communication = col\_double(),  
 .. review\_scores\_location = col\_double(),  
 .. review\_scores\_value = col\_double(),  
 .. license = col\_character(),  
 .. instant\_bookable = col\_logical(),  
 .. calculated\_host\_listings\_count = col\_double(),  
 .. calculated\_host\_listings\_count\_entire\_homes = col\_double(),  
 .. calculated\_host\_listings\_count\_private\_rooms = col\_double(),  
 .. calculated\_host\_listings\_count\_shared\_rooms = col\_double(),  
 .. reviews\_per\_month = col\_double()  
 .. )  
 - attr(\*, "problems")=<externalptr>

En el archivo encontramos cuatro tipos de datos distintos:

* double
* character
* logical
* date

Los datos nos hablan de todos los apartados relacionados con el alojamiento ofrecido, la información del host, el precio medio, la disponibilidad, imagenes de los alojamientos e información de las reseñas.

## Definición de variables

Vamos a definir las variables que contienen los datos

|  |  |
| --- | --- |
| id | Identificador único del alojamiento en Airbnb |
| listing\_url | Link al portal del alojamiento |
| scrape\_id | Identificador de scrapping de los datos |
| last\_scraped | Hora del scrapping |
| source | Fuente, de la entrada, si es reciente o antigua. |
| name | Nombre del alojamiento |
| description | Descripción del alojamiento |
| neighborhood\_overview | Descripción del barrio del host |
| picture\_url | Link de las fotografías del alojamiento |
| host\_id | Identificador del host del alojamiento |
| host\_url | Link al portal del perfil del host |
| host\_name | Nombre del host |
| host\_since | Fecha de registro como host |
| host\_location | Localización del host |
| host\_about | Descripción del host sobre si mismo. |
| host\_response\_time | Tiempo promedio de respuesta del host |
| host\_response\_rate | Porcentaje de respuestas realizadas por el host. |
| host\_acceptance\_rate | Porcentaje de solicitudes de alojamiento aceptadas por el host. |
| host\_is\_superhost | Estatus otorgado por Airbnb por ser un host con ciertas características positivas según la plataforma |
| host\_picture\_url | Link de la foto de perfil del host |
| host\_neighbourhood | Barrio donde reside el host |
| host\_listings\_count | Número de alojamientos listados por el host (cuenta manual) |
| host\_total\_listings\_count | Número de alojamientos listados por el host (cuenta segun Airbnb) |
| host\_verifications | Métodos por los cuales el host se ha verificado |
| host\_has\_profile\_pic | Boolean que indica si el host se ha verificado |
| host\_identity\_verified | Boolean que indica si la identidad se ha verificado |
| neighbourhood\_group\_cleansed | Barrio según coordenadas. |
| latitude | Coordenada de latitud del alojamiento |
| longitude | Coordenada de longitud del alojamiento |
| property\_type | Tipo de propiedad del alojamiento |
| room\_type | Tipo de habitación, si es compartida o individual. |
| accommodates | Capacidad de personas |
| bathrooms | Número de banos |
| bedrooms | Número de dormitorios |
| beds | Número de camas |
| price | Precio diario para alojarse |
| minimum\_nights | Noches mínimas que hay alojarse para reservar |
| maximum\_nights | Noches máximas que se pueden reservar |
| has\_availability | Indica si hay disponibilidad. |
| availability\_30 | Disponibilidad en los proximos 30 dias. |
| availability\_60 | Disponibilidad en los proximos 60 dias. |
| availability\_90 | Disponibilidad en los proximos 90 dias. |
| availability\_365 | Disponibilidad en los proximos 365 dias. |
| number\_of\_reviews | Número de reviews que ha recibido el alojamiento. |
| number\_of\_reviews\_ltm | Número de reviews que ha recibido el alojamiento en los últimos 12 meses. |
| number\_of\_reviews\_l30d | Número de reviews que ha recibido el alojamiento en los últimos 30 dias. |
| first\_review | Fecha de la primera review del alojamiento. |
| last\_review | Fecha de la última review del alojamiento |
| review\_scores\_rating | Nota media de las reviews en general |
| review\_scores\_accuracy | Nota media de las reviews en precisión |
| review\_scores\_cleanliness | Nota media de las reviews en limpieza |
| review\_scores\_checkin | Nota media de las reviews en el proceso de checkin |
| review\_scores\_communication | Nota media de las reviews en comunicación |
| review\_scores\_location | Nota media de las reviews en localización |
| review\_scores\_value | Nota media de las reviews en valor (calidad / precio) |
| license | Licencia del alojamiento |
| instant\_bookable | Indica si el alojamiento se reserva automáticamente o se necesita la aprobación del host. |

## Muestra de los datos

A continuación, una muestra de los datos obtenidos:

head(df)

# A tibble: 6 × 75  
 id listing\_url scrape\_id last\_scraped source name description  
 <dbl> <chr> <dbl> <date> <chr> <chr> <chr>   
1 18674 https://www.airbnb.com… 2.02e13 2023-09-06 city … Rent… 110m2 apar…  
2 23197 https://www.airbnb.com… 2.02e13 2023-09-06 city … Rent… Beautiful …  
3 117010 https://www.airbnb.com… 2.02e13 2023-09-06 previ… Rent… Have an au…  
4 32711 https://www.airbnb.com… 2.02e13 2023-09-06 city … Rent… A lovely t…  
5 118228 https://www.airbnb.com… 2.02e13 2023-09-06 previ… Rent… Modern 100…  
6 128463 https://www.airbnb.com… 2.02e13 2023-09-06 city … Rent… My House i…  
# ℹ 68 more variables: neighborhood\_overview <chr>, picture\_url <chr>,  
# host\_id <dbl>, host\_url <chr>, host\_name <chr>, host\_since <date>,  
# host\_location <chr>, host\_about <chr>, host\_response\_time <chr>,  
# host\_response\_rate <chr>, host\_acceptance\_rate <chr>,  
# host\_is\_superhost <lgl>, host\_thumbnail\_url <chr>, host\_picture\_url <chr>,  
# host\_neighbourhood <chr>, host\_listings\_count <dbl>,  
# host\_total\_listings\_count <dbl>, host\_verifications <chr>, …

Vamos a ver un análisis rápido de los datos:

summary(df)

id listing\_url scrape\_id   
 Min. :1.867e+04 Length:18086 Min. :2.023e+13   
 1st Qu.:2.172e+07 Class :character 1st Qu.:2.023e+13   
 Median :4.435e+07 Mode :character Median :2.023e+13   
 Mean :2.997e+17 Mean :2.023e+13   
 3rd Qu.:7.450e+17 3rd Qu.:2.023e+13   
 Max. :9.740e+17 Max. :2.023e+13   
   
 last\_scraped source name description   
 Min. :2023-09-06 Length:18086 Length:18086 Length:18086   
 1st Qu.:2023-09-06 Class :character Class :character Class :character   
 Median :2023-09-06 Mode :character Mode :character Mode :character   
 Mean :2023-09-06   
 3rd Qu.:2023-09-06   
 Max. :2023-09-06   
   
 neighborhood\_overview picture\_url host\_id   
 Length:18086 Length:18086 Min. : 3073   
 Class :character Class :character 1st Qu.: 9919300   
 Mode :character Mode :character Median : 96299106   
 Mean :166184865   
 3rd Qu.:310348791   
 Max. :535400790   
   
 host\_url host\_name host\_since host\_location   
 Length:18086 Length:18086 Min. :2008-09-19 Length:18086   
 Class :character Class :character 1st Qu.:2013-11-10 Class :character   
 Mode :character Mode :character Median :2016-09-22 Mode :character   
 Mean :2016-12-27   
 3rd Qu.:2019-11-18   
 Max. :2023-09-04   
 NA's :2   
 host\_about host\_response\_time host\_response\_rate host\_acceptance\_rate  
 Length:18086 Length:18086 Length:18086 Length:18086   
 Class :character Class :character Class :character Class :character   
 Mode :character Mode :character Mode :character Mode :character   
   
   
   
   
 host\_is\_superhost host\_thumbnail\_url host\_picture\_url host\_neighbourhood  
 Mode :logical Length:18086 Length:18086 Length:18086   
 FALSE:14580 Class :character Class :character Class :character   
 TRUE :3087 Mode :character Mode :character Mode :character   
 NA's :419   
   
   
   
 host\_listings\_count host\_total\_listings\_count host\_verifications  
 Min. : 1.00 Min. : 1.00 Length:18086   
 1st Qu.: 2.00 1st Qu.: 2.00 Class :character   
 Median : 6.00 Median : 9.00 Mode :character   
 Mean : 41.13 Mean : 53.55   
 3rd Qu.: 34.00 3rd Qu.: 48.00   
 Max. :786.00 Max. :1853.00   
 NA's :2 NA's :2   
 host\_has\_profile\_pic host\_identity\_verified neighbourhood   
 Mode :logical Mode :logical Length:18086   
 FALSE:330 FALSE:1471 Class :character   
 TRUE :17754 TRUE :16613 Mode :character   
 NA's :2 NA's :2   
   
   
   
 neighbourhood\_cleansed neighbourhood\_group\_cleansed latitude   
 Length:18086 Length:18086 Min. :41.35   
 Class :character Class :character 1st Qu.:41.38   
 Mode :character Mode :character Median :41.39   
 Mean :41.39   
 3rd Qu.:41.40   
 Max. :41.46   
   
 longitude property\_type room\_type accommodates   
 Min. :2.092 Length:18086 Length:18086 Min. : 1.000   
 1st Qu.:2.157 Class :character Class :character 1st Qu.: 2.000   
 Median :2.168 Mode :character Mode :character Median : 3.000   
 Mean :2.167 Mean : 3.365   
 3rd Qu.:2.177 3rd Qu.: 4.000   
 Max. :2.228 Max. :16.000   
   
 bathrooms bathrooms\_text bedrooms beds   
 Mode:logical Length:18086 Min. : 1.000 Min. : 1.000   
 NA's:18086 Class :character 1st Qu.: 1.000 1st Qu.: 1.000   
 Mode :character Median : 2.000 Median : 2.000   
 Mean : 2.066 Mean : 2.358   
 3rd Qu.: 3.000 3rd Qu.: 3.000   
 Max. :12.000 Max. :30.000   
 NA's :6360 NA's :314   
 amenities price minimum\_nights maximum\_nights   
 Length:18086 Length:18086 Min. : 1.00 Min. : 1.0   
 Class :character Class :character 1st Qu.: 1.00 1st Qu.: 180.2   
 Mode :character Mode :character Median : 3.00 Median : 365.0   
 Mean : 14.76 Mean : 574.4   
 3rd Qu.: 31.00 3rd Qu.:1125.0   
 Max. :1125.00 Max. :3000.0   
   
 minimum\_minimum\_nights maximum\_minimum\_nights minimum\_maximum\_nights  
 Min. : 1.00 Min. : 1.00 Min. :1.000e+00   
 1st Qu.: 1.00 1st Qu.: 3.00 1st Qu.:3.000e+02   
 Median : 2.00 Median : 5.00 Median :3.650e+02   
 Mean : 14.54 Mean : 18.89 Mean :2.381e+05   
 3rd Qu.: 31.00 3rd Qu.: 31.00 3rd Qu.:1.125e+03   
 Max. :1125.00 Max. :2705.00 Max. :2.147e+09   
   
 maximum\_maximum\_nights minimum\_nights\_avg\_ntm maximum\_nights\_avg\_ntm  
 Min. :1.000e+00 Min. : 1.00 Min. :1.000e+00   
 1st Qu.:3.330e+02 1st Qu.: 2.00 1st Qu.:3.300e+02   
 Median :9.990e+02 Median : 3.70 Median :4.070e+02   
 Mean :2.382e+05 Mean : 17.34 Mean :2.381e+05   
 3rd Qu.:1.125e+03 3rd Qu.: 31.00 3rd Qu.:1.125e+03   
 Max. :2.147e+09 Max. :1125.00 Max. :2.147e+09   
   
 calendar\_updated has\_availability availability\_30 availability\_60  
 Mode:logical Mode :logical Min. : 0.000 Min. : 0.00   
 NA's:18086 FALSE:1188 1st Qu.: 0.000 1st Qu.: 0.00   
 TRUE :16898 Median : 4.000 Median :13.00   
 Mean : 7.718 Mean :19.37   
 3rd Qu.:12.000 3rd Qu.:34.00   
 Max. :30.000 Max. :60.00   
   
 availability\_90 availability\_365 calendar\_last\_scraped number\_of\_reviews  
 Min. : 0.00 Min. : 0.0 Min. :2023-09-06 Min. : 0.00   
 1st Qu.: 1.00 1st Qu.: 47.0 1st Qu.:2023-09-06 1st Qu.: 1.00   
 Median :35.00 Median :175.0 Median :2023-09-06 Median : 6.00   
 Mean :36.43 Mean :171.8 Mean :2023-09-06 Mean : 42.22   
 3rd Qu.:62.00 3rd Qu.:302.0 3rd Qu.:2023-09-06 3rd Qu.: 42.00   
 Max. :90.00 Max. :365.0 Max. :2023-09-06 Max. :1817.00   
   
 number\_of\_reviews\_ltm number\_of\_reviews\_l30d first\_review   
 Min. : 0.00 Min. : 0.0000 Min. :2010-10-03   
 1st Qu.: 0.00 1st Qu.: 0.0000 1st Qu.:2017-08-16   
 Median : 2.00 Median : 0.0000 Median :2020-06-28   
 Mean : 11.44 Mean : 0.9239 Mean :2019-12-27   
 3rd Qu.: 15.00 3rd Qu.: 1.0000 3rd Qu.:2022-09-25   
 Max. :836.00 Max. :91.0000 Max. :2023-09-05   
 NA's :4466   
 last\_review review\_scores\_rating review\_scores\_accuracy  
 Min. :2011-06-23 Min. :0.00 Min. :0.000   
 1st Qu.:2023-01-31 1st Qu.:4.41 1st Qu.:4.500   
 Median :2023-08-02 Median :4.68 Median :4.750   
 Mean :2022-10-26 Mean :4.54 Mean :4.621   
 3rd Qu.:2023-08-22 3rd Qu.:4.90 3rd Qu.:4.930   
 Max. :2023-09-06 Max. :5.00 Max. :5.000   
 NA's :4466 NA's :4466 NA's :4549   
 review\_scores\_cleanliness review\_scores\_checkin review\_scores\_communication  
 Min. :0.000 Min. :0.000 Min. :0.000   
 1st Qu.:4.450 1st Qu.:4.640 1st Qu.:4.640   
 Median :4.720 Median :4.850 Median :4.850   
 Mean :4.586 Mean :4.711 Mean :4.709   
 3rd Qu.:4.920 3rd Qu.:5.000 3rd Qu.:5.000   
 Max. :5.000 Max. :5.000 Max. :5.000   
 NA's :4548 NA's :4553 NA's :4547   
 review\_scores\_location review\_scores\_value license instant\_bookable  
 Min. :0.000 Min. :0.000 Length:18086 Mode :logical   
 1st Qu.:4.670 1st Qu.:4.280 Class :character FALSE:11248   
 Median :4.830 Median :4.560 Mode :character TRUE :6838   
 Mean :4.737 Mean :4.444   
 3rd Qu.:5.000 3rd Qu.:4.780   
 Max. :5.000 Max. :5.000   
 NA's :4552 NA's :4553   
 calculated\_host\_listings\_count calculated\_host\_listings\_count\_entire\_homes  
 Min. : 1.00 Min. : 0.00   
 1st Qu.: 1.00 1st Qu.: 0.00   
 Median : 5.00 Median : 2.00   
 Mean : 31.63 Mean : 23.87   
 3rd Qu.: 28.00 3rd Qu.: 19.00   
 Max. :294.00 Max. :294.00   
   
 calculated\_host\_listings\_count\_private\_rooms  
 Min. : 0.000   
 1st Qu.: 0.000   
 Median : 0.000   
 Mean : 7.602   
 3rd Qu.: 2.000   
 Max. :233.000   
   
 calculated\_host\_listings\_count\_shared\_rooms reviews\_per\_month  
 Min. : 0.00000 Min. : 0.010   
 1st Qu.: 0.00000 1st Qu.: 0.240   
 Median : 0.00000 Median : 0.850   
 Mean : 0.08332 Mean : 1.436   
 3rd Qu.: 0.00000 3rd Qu.: 2.100   
 Max. :12.00000 Max. :55.020   
 NA's :4466

# Objetivo del proyecto

El objetivo con el análisis de datos es crear un modelo el cual enseñe a actuales o futuros propietarios de alojamientos de airbnb a obtener los máximos ingresos posibles teniendo en cuenta todas las características en del alojamiento y su perfil en Airbnb.

La variable a predecir es “price”, la cual es el precio de reservar cada alojamiento en su fecha de disponiblidad mas reciente.

El modelo ideal será en el que haya menos residuo entre los valores reales y los valores predichos.

## Modelo escogido

El modelo escogido es una regresión lineal, que es de aprendizaje supervisado y continuo. Hemos tenido en cuenta otros posibles modelos y hemos llegado a las siguientes conclusiones:

| Modelo | Pros | Contras |
| --- | --- | --- |
| Regresión lineal | * Es simple y fácil de entender, los resultados pueden ser visualizados rapidamente. * Los coeficientes de regresión pueden interpretarse directamente. | * Se basa en supuestos estrictos que muchas veces no se cumplen en datos reales. * Es sensible a valores atípicos. |
| Regresión Ridge | * Controla el sobreajuste, penalizando a los coeficientes grandes. * Menos sensible a valores atípicos. | * No realiza selección de variables, todos los predictores se mantienen en el modelo. * Requiere la selección del parámetros de regularización. |
| Regresión Lasso | * Es capaz de realizar selección de variables reduciendo algunos coeficientes a cero. * Útil cuando se tienen muchas variables y se desea simplificar el modelo. | * Puede ser inestable en la presencia de predictores altamente correlacionados. * También requiere la selección cuidadosa del parámetro de regularización. |

## Programa y paquetes seleccionados: Decisión y razones

R es un lenguaje de programación muy utilizado en estadística y aprendizaje automático debido a su gran cantidad de paquetes y funciones para análisis de datos y modelado estadístico. Su popularidad entre los estadísticos y científicos de datos se debe a la facilidad con la que se pueden manipular datos, realizar cálculos estadísticos y generar gráficos avanzados. Para esta funcionalidad concreta, en los ultimos años, han destacado dos grupos de liberías que también vamos a utilizar en nuestro modelo, como son Tidyverse y Tidymodels.

El tidyverse es una colección de paquetes de R diseñados para la ciencia de datos que comparten una filosofía subyacente de diseño y gramática. Algunos de los paquetes más destacados del tidyverse incluyen:

* dplyr y tidyr para la manipulación de datos.
* Ggplot para la creación de visualizaciones avanzadas.

Tidymodels tiene el objetivo de proporcionar una gramática coherente y fácil de usar para la modelización predictiva en R. Algunos paquetes son:

* rsample que contiene funciones para la separación de sets de un dataframe
* yardstick: Incluye métricas para medir el rendimiento de los modelos.

Ambos grupos de librerias tienen finalidades similares como el uso de una grámatica consistente, el formato con pipes y en general que sea legible y simple.

# Limpieza de datos

En primer lugar vamos a eliminar aquellas variables que contienen datos que no vamos a utilizar por alguno de los siguientes motivos:

* Corresponden a links de cada una de las entradas
* Id’s de elementos concretos que no vamos a utilizar
* Descripciones y otros elementos strings
* Variables cubiertas por otras las cuales estandarizan el formato

df <- df |>  
 select(c(id,  
 host\_id,  
 host\_since,  
 host\_response\_time,  
 host\_response\_rate,  
 host\_acceptance\_rate,  
 host\_is\_superhost,  
 host\_verifications,  
 host\_has\_profile\_pic,  
 host\_identity\_verified,  
 calculated\_host\_listings\_count,  
 calculated\_host\_listings\_count\_entire\_homes,  
 calculated\_host\_listings\_count\_private\_rooms,  
 calculated\_host\_listings\_count\_shared\_rooms,  
 neighbourhood\_group\_cleansed,  
 latitude,  
 longitude,  
 room\_type,  
 accommodates,  
 bathrooms\_text,  
 bedrooms,  
 beds,  
 price,  
 minimum\_nights,  
 maximum\_nights,  
 has\_availability,  
 availability\_30,  
 availability\_60,  
 availability\_90,  
 availability\_365,  
 instant\_bookable,  
 number\_of\_reviews,  
 number\_of\_reviews\_l30d,  
 review\_scores\_rating,  
 review\_scores\_accuracy,  
 review\_scores\_cleanliness,  
 review\_scores\_checkin,  
 review\_scores\_communication,  
 review\_scores\_location,  
 review\_scores\_value,  
 ))

## Ajustes de formato

**host\_response\_time**

Nos encontramos con variables categoricas en diferentes medidas de tiempo. Sería conveniente ajustar todas estas pora que estén en la misma medida de tiempo.

unique(df$host\_response\_time)

[1] "within an hour" "within a few hours" "within a day"   
[4] "N/A" "a few days or more" NA

Encontramos seis valores. A continuación listamos las diferentes categorías y su nuevo valor: - “within an hour” -> <1h - “within a few hours” -> ~12h - “within a day -> <24h -”a few days or more” -> >48h - “N/A” -> Unificar con NA

df$host\_response\_time <- ifelse(df$host\_response\_time == "within an hour", "<1h",  
 ifelse(df$host\_response\_time == "within a few hours", "~12h",  
 ifelse(df$host\_response\_time == "within a day", "~24h",  
 ifelse(df$host\_response\_time == "a few days or more", ">48h",  
 ifelse(df$host\_response\_time == "N/A", NA, df$host\_response\_time)))))

**bathrooms\_text**

Esta variable incluye el valor númerico de baños y además nos indica si estos son compartidos o no. Para la claridad en los datos sería conveniente crear dos variables de esta:

1. Una variable número con el número de baños.
2. Un Boolean indicando si estos son compartidos o no.

unique(df$bathrooms\_text)

[1] "2 baths" "1.5 baths" "2.5 baths"   
 [4] "3 baths" "1 shared bath" "1 bath"   
 [7] "1 private bath" "3.5 baths" "4 baths"   
[10] "1.5 shared baths" NA "2 shared baths"   
[13] "2.5 shared baths" "5.5 baths" "7.5 baths"   
[16] "4.5 baths" "6 baths" "0 shared baths"   
[19] "Half-bath" "Private half-bath" "0 baths"   
[22] "5 baths" "8 baths" "3 shared baths"   
[25] "8 shared baths" "4 shared baths" "Shared half-bath"   
[28] "5 shared baths" "3.5 shared baths" "6 shared baths"   
[31] "5.5 shared baths" "10 baths" "10 shared baths"   
[34] "6.5 baths" "4.5 shared baths"

Vemos que casi todas las categorías tienen números. A excepción de tres de ellas. Vamos a añadirselo para compartir la estructura:

df$bathrooms\_text <- ifelse(df$bathrooms\_text == "Shared half-bath", "0.5 Shared bath",  
 ifelse(df$bathrooms\_text == "Private half-bath", "0.5 bath",  
 ifelse(df$bathrooms\_text == "Half-bath", "0.5 bath", df$bathrooms\_text)))

A continuación creamos la columna boolean la cual nos indica si un baño es compartido o no (Asumimos que si no es indicado, es privado).

df$shared\_bathrooms <- grepl("shared|Shared", df$bathrooms\_text)

Seguidamente creamos una nueva columna que extraiga el número de baños del valor bathrooms\_text

df$n\_bathrooms <- sapply(strsplit(df$bathrooms\_text, " "), function(x) as.numeric(x[1]))

Recolocamos las columnas en la posición de bathrooms\_text y eliminamos esta.

bathrooms\_text\_position <- which(colnames(df) == "bathrooms\_text")  
  
df <- df %>%  
 relocate(all\_of(c("n\_bathrooms", "shared\_bathrooms")), .before = bathrooms\_text\_position)

Warning: Using an external vector in selections was deprecated in tidyselect 1.1.0.  
ℹ Please use `all\_of()` or `any\_of()` instead.  
 # Was:  
 data %>% select(bathrooms\_text\_position)  
  
 # Now:  
 data %>% select(all\_of(bathrooms\_text\_position))  
  
See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.

df <- df %>%  
 select(-bathrooms\_text)  
  
rm(bathrooms\_text\_position)

**host\_verifications**

En esta variable podemos ver en formato lista, los medios por los cuales se ha verificado un host. Para que los datos sean mas accesibles podemos crear una variable de tipo boolean por cada tipo de medio posible de verificación.

unique(df$host\_verifications)

[1] "['email', 'phone']" "['email', 'phone', 'work\_email']"  
[3] "['phone', 'work\_email']" "['phone']"   
[5] "['email']" "None"   
[7] "[]" "['work\_email']"

Encontramos tres tipos de verificación:

* “email”
* “phone”
* “work\_email”

Creamos las tres variables y añadimos true si ese método está incluido en la columna host\_verifications.

df <- df %>%  
 mutate(verification\_email = str\_detect(df$host\_verifications, "email"),  
 verification\_phone = str\_detect(df$host\_verifications, "phone"),  
 verification\_work\_email = str\_detect(df$host\_verifications, "work\_email"))  
  
host\_verification\_text\_position <- which(colnames(df) == "host\_verifications")  
  
df <- df %>%  
 relocate(all\_of(c("verification\_email", "verification\_phone", "verification\_work\_email")), .before = host\_verification\_text\_position)

Warning: Using an external vector in selections was deprecated in tidyselect 1.1.0.  
ℹ Please use `all\_of()` or `any\_of()` instead.  
 # Was:  
 data %>% select(host\_verification\_text\_position)  
  
 # Now:  
 data %>% select(all\_of(host\_verification\_text\_position))  
  
See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.

df <- df %>%  
 select(-host\_verifications)  
  
rm(host\_verification\_text\_position)

**price**

Podemos convertir price en una variable numérica para poderla usar en visualizaciones y otros tipos de funciones de tipo continuo:

df$price <- as.numeric(gsub("\\$", "", df$price))

Warning: NAs introduced by coercion

**review\_score\_rating**

El valor mínimo que puede tener un alojamiento es 1. En algunos campos pone 0 a causa de que el resto de valores son NA. Vamos a modificar estos valores a NA.

df$review\_scores\_rating <- ifelse(df$review\_scores\_rating == 0, NA, df$review\_scores\_rating)

**host\_response\_rate y host\_acceptance\_rate**

Cambiamos ambas variables de tipo character con porcentaje a tipo numérico double:

df$host\_response\_rate <- as.numeric(sub("%", "", df$host\_response\_rate))/100

Warning: NAs introduced by coercion

df$host\_acceptance\_rate <- as.numeric(sub("%", "", df$host\_acceptance\_rate))/100

Warning: NAs introduced by coercion

**Conversión de campos con tags en factores**

Cambiamos todas las columnas de tipos strings que tengan etiquetas a tipo factor para que su uso sea mas sencillo.

df <- df %>%  
 mutate(  
 host\_response\_time = factor(host\_response\_time),  
 neighbourhood\_group\_cleansed = factor(neighbourhood\_group\_cleansed),  
 room\_type = factor(room\_type)  
 )

## Tratamiento de NA’s

Comprobamos cuantos NA’s faltan en cada columna para ver que tratamiento hacer con ellos.

na\_analysis <- df %>% summarize(across(everything(), ~sum(is.na(.))))  
na\_analysis

# A tibble: 1 × 43  
 id host\_id host\_since host\_response\_time host\_response\_rate  
 <int> <int> <int> <int> <int>  
1 0 0 2 2963 2963  
# ℹ 38 more variables: host\_acceptance\_rate <int>, host\_is\_superhost <int>,  
# verification\_email <int>, verification\_phone <int>,  
# verification\_work\_email <int>, host\_has\_profile\_pic <int>,  
# host\_identity\_verified <int>, calculated\_host\_listings\_count <int>,  
# calculated\_host\_listings\_count\_entire\_homes <int>,  
# calculated\_host\_listings\_count\_private\_rooms <int>,  
# calculated\_host\_listings\_count\_shared\_rooms <int>, …

**Columnas numéricas**

En primer lugar vamos a promediar todas las columnas númericas:

df <- df |>  
 mutate(across(where(is.numeric), ~if (any(is.na(.))) floor(replace\_na(., mean(., na.rm = TRUE))) else .))

**host\_response\_time**

Vamos a hacer una tablaq para ver la distribución de cada categoría:

frecuencias\_host\_response\_time <- table(df$host\_response\_time)  
porcentajes\_host\_response\_time <- prop.table(frecuencias\_host\_response\_time) \* 100  
print(porcentajes\_host\_response\_time)

<1h >48h ~12h ~24h   
71.936785 1.937446 15.823580 10.302189

Vemos como la gran mayoría responde en menos de una hora, al no ser un componente mayor esta variable vamos a justar los NA’s como menores de una hora.

df$host\_response\_time[is.na(df$host\_response\_time)] <- "<1h"

**host\_is\_superhost**

Como no tenemos información asumimos que los valores NA son false

df$host\_is\_superhost[is.na(df$host\_is\_superhost)] <- FALSE

**host\_has\_profile\_pic**

Hacemos la misma gestión en este caso

df$host\_has\_profile\_pic[is.na(df$host\_has\_profile\_pic)] <- FALSE

**host\_identity\_verified**

Volvemos a convertir los valores NA que tenemos en FALSE

df$host\_identity\_verified[is.na(df$host\_identity\_verified)] <- FALSE

**host\_since**

Vamos a usar el valor promedio para sustituir los NA de la fecha de ingreso de los host

mean\_host\_since <- mean(df$host\_since, na.rm=TRUE)  
df$host\_since[is.na(df$host\_since)] <- mean\_host\_since

# Análisis descriptivo de los datos

Vamos a hacer un análisis descriptivo de nuestros datos para poder tendencias y patrones para poder tomar decisiones sobre el modelo mas adelante.

## Summary

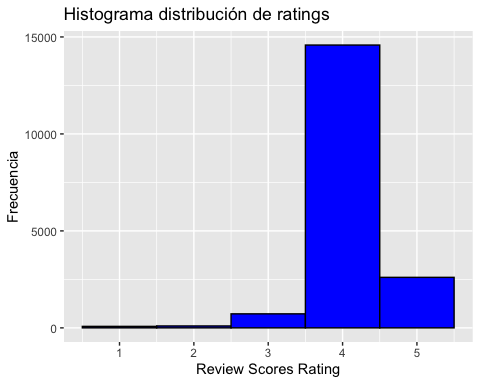
Vamos a ver como quedaría el summary después de haber limpiado los datos.

summary(df)

id host\_id host\_since   
 Min. :1.867e+04 Min. : 3073 Min. :2008-09-19   
 1st Qu.:2.172e+07 1st Qu.: 9919300 1st Qu.:2013-11-10   
 Median :4.435e+07 Median : 96299106 Median :2016-09-22   
 Mean :2.997e+17 Mean :166184865 Mean :2016-12-27   
 3rd Qu.:7.450e+17 3rd Qu.:310348791 3rd Qu.:2019-11-17   
 Max. :9.740e+17 Max. :535400790 Max. :2023-09-04   
   
 host\_response\_time host\_response\_rate host\_acceptance\_rate host\_is\_superhost  
 <1h :13842 Min. :0.0000 Min. :0.0000 Mode :logical   
 >48h: 293 1st Qu.:0.0000 1st Qu.:0.0000 FALSE:14999   
 ~12h: 2393 Median :0.0000 Median :0.0000 TRUE :3087   
 ~24h: 1558 Mean :0.4511 Mean :0.2769   
 3rd Qu.:1.0000 3rd Qu.:1.0000   
 Max. :1.0000 Max. :1.0000   
   
 verification\_email verification\_phone verification\_work\_email  
 Mode :logical Mode :logical Mode :logical   
 FALSE:1264 FALSE:43 FALSE:15153   
 TRUE :16822 TRUE :18043 TRUE :2933   
   
   
   
   
 host\_has\_profile\_pic host\_identity\_verified calculated\_host\_listings\_count  
 Mode :logical Mode :logical Min. : 1.00   
 FALSE:332 FALSE:1473 1st Qu.: 1.00   
 TRUE :17754 TRUE :16613 Median : 5.00   
 Mean : 31.63   
 3rd Qu.: 28.00   
 Max. :294.00   
   
 calculated\_host\_listings\_count\_entire\_homes  
 Min. : 0.00   
 1st Qu.: 0.00   
 Median : 2.00   
 Mean : 23.87   
 3rd Qu.: 19.00   
 Max. :294.00   
   
 calculated\_host\_listings\_count\_private\_rooms  
 Min. : 0.000   
 1st Qu.: 0.000   
 Median : 0.000   
 Mean : 7.602   
 3rd Qu.: 2.000   
 Max. :233.000   
   
 calculated\_host\_listings\_count\_shared\_rooms neighbourhood\_group\_cleansed  
 Min. : 0.00000 Eixample :6469   
 1st Qu.: 0.00000 Ciutat Vella :4218   
 Median : 0.00000 Sants-Montjuïc :1883   
 Mean : 0.08332 Sant Martí :1662   
 3rd Qu.: 0.00000 Gràcia :1591   
 Max. :12.00000 Sarrià-Sant Gervasi: 890   
 (Other) :1373   
 latitude longitude room\_type accommodates   
 Min. :41.35 Min. :2.092 Entire home/apt:10622 Min. : 1.000   
 1st Qu.:41.38 1st Qu.:2.157 Hotel room : 134 1st Qu.: 2.000   
 Median :41.39 Median :2.168 Private room : 7173 Median : 3.000   
 Mean :41.39 Mean :2.167 Shared room : 157 Mean : 3.365   
 3rd Qu.:41.40 3rd Qu.:2.177 3rd Qu.: 4.000   
 Max. :41.46 Max. :2.228 Max. :16.000   
   
 n\_bathrooms shared\_bathrooms bedrooms beds   
 Min. : 0.000 Mode :logical Min. : 1.000 Min. : 1.000   
 1st Qu.: 1.000 FALSE:13774 1st Qu.: 2.000 1st Qu.: 1.000   
 Median : 1.000 TRUE :4312 Median : 2.000 Median : 2.000   
 Mean : 1.309 Mean : 2.043 Mean : 2.351   
 3rd Qu.: 2.000 3rd Qu.: 2.000 3rd Qu.: 3.000   
 Max. :10.000 Max. :12.000 Max. :30.000   
   
 price minimum\_nights maximum\_nights has\_availability  
 Min. : 8.0 Min. : 1.00 Min. : 1.0 Mode :logical   
 1st Qu.: 52.0 1st Qu.: 1.00 1st Qu.: 180.2 FALSE:1188   
 Median :100.0 Median : 3.00 Median : 365.0 TRUE :16898   
 Mean :135.1 Mean : 14.76 Mean : 574.4   
 3rd Qu.:182.0 3rd Qu.: 31.00 3rd Qu.:1125.0   
 Max. :999.0 Max. :1125.00 Max. :3000.0   
   
 availability\_30 availability\_60 availability\_90 availability\_365  
 Min. : 0.000 Min. : 0.00 Min. : 0.00 Min. : 0.0   
 1st Qu.: 0.000 1st Qu.: 0.00 1st Qu.: 1.00 1st Qu.: 47.0   
 Median : 4.000 Median :13.00 Median :35.00 Median :175.0   
 Mean : 7.718 Mean :19.37 Mean :36.43 Mean :171.8   
 3rd Qu.:12.000 3rd Qu.:34.00 3rd Qu.:62.00 3rd Qu.:302.0   
 Max. :30.000 Max. :60.00 Max. :90.00 Max. :365.0   
   
 instant\_bookable number\_of\_reviews number\_of\_reviews\_l30d review\_scores\_rating  
 Mode :logical Min. : 0.00 Min. : 0.0000 Min. :1.00   
 FALSE:11248 1st Qu.: 1.00 1st Qu.: 0.0000 1st Qu.:4.00   
 TRUE :6838 Median : 6.00 Median : 0.0000 Median :4.00   
 Mean : 42.22 Mean : 0.9239 Mean :4.08   
 3rd Qu.: 42.00 3rd Qu.: 1.0000 3rd Qu.:4.00   
 Max. :1817.00 Max. :91.0000 Max. :5.00   
   
 review\_scores\_accuracy review\_scores\_cleanliness review\_scores\_checkin  
 Min. :0.0 Min. :0.000 Min. :0.000   
 1st Qu.:4.0 1st Qu.:4.000 1st Qu.:4.000   
 Median :4.0 Median :4.000 Median :4.000   
 Mean :4.1 Mean :4.085 Mean :4.165   
 3rd Qu.:4.0 3rd Qu.:4.000 3rd Qu.:4.000   
 Max. :5.0 Max. :5.000 Max. :5.000   
   
 review\_scores\_communication review\_scores\_location review\_scores\_value  
 Min. :0.000 Min. :0.000 Min. :0.000   
 1st Qu.:4.000 1st Qu.:4.000 1st Qu.:4.000   
 Median :4.000 Median :4.000 Median :4.000   
 Mean :4.165 Mean :4.167 Mean :4.009   
 3rd Qu.:4.000 3rd Qu.:4.000 3rd Qu.:4.000   
 Max. :5.000 Max. :5.000 Max. :5.000

## Distribución de las notas de review

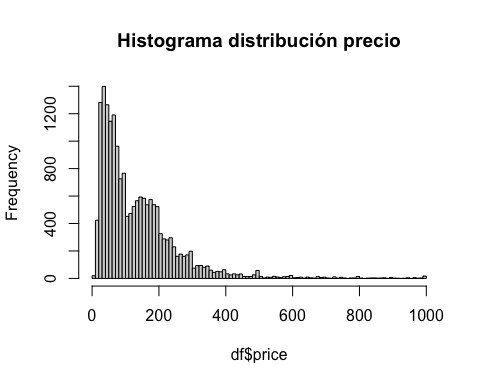
ggplot(df, aes(x = review\_scores\_rating)) +  
 geom\_histogram(binwidth = 1, fill = "blue", color = "black") +  
 ggtitle("Histograma distribución de ratings") +  
 xlab("Review Scores Rating") +  
 ylab("Frecuencia")



En el gráfico podemos ver como la gran mayoría de reseñas tienen un nivel de entre 4 y 5, siendo 4 con mucha diferencia la moda. Vemos como las notas entre 1 y 3 son mínimas.

## Distribución del precio por noche

hist(df$price, main = "Histograma distribución precio", breaks=100)

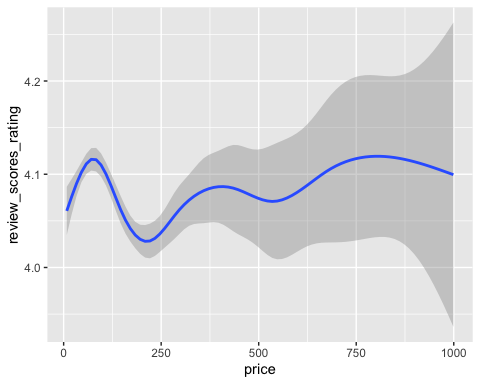


El rango de precios con mayor frecuencia esta entre 0 y 100, por otro lado existen pocas observaciones con precios superiores a 400. La distribución de los precios es asimétrica positiva (sesgada a la derecha), con una cola larga que se extiende hacia los precios más altos lo que implica que las frecuencias de precios disminuyen rápidamente a medida que aumenta el precio.

## Distribución entre precio y reviews

ggplot(df, aes(price, review\_scores\_rating)) +   
 geom\_smooth()

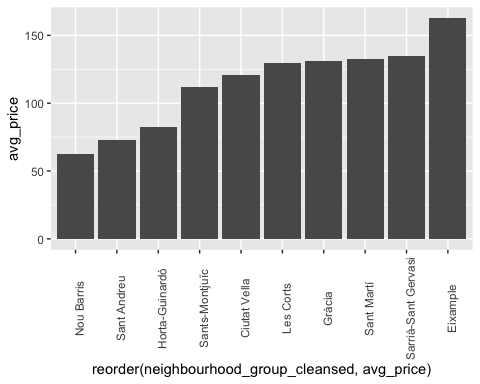
`geom\_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'



Existe una tendencia que muestra que a medida que el precio aumenta, las calificaciones de revisión tienden a mejorar, especialmente después del punto de precio de 750. El intervalo de confianza se amplía a medida que el precio aumenta, lo que sugiere que hay más variabilidad en las calificaciones de revisión para productos o servicios más caros. La mayoría de los datos parece estar concentrada en el rango de precios más bajo.

## Media de precio por barrio

df\_precio\_barrio <- df %>% select(price, neighbourhood\_group\_cleansed) %>% group\_by(neighbourhood\_group\_cleansed) %>% summarize(avg\_price = mean(price, na.rm = TRUE))  
  
ggplot(df\_precio\_barrio, aes(reorder(neighbourhood\_group\_cleansed, avg\_price), avg\_price)) +  
 geom\_col() +  
 theme(axis.text.x = element\_text(angle = 90))



Hay una variabilidad notable en los precios promedio entre los diferentes barrios. Eixample parece ser el barrio con el precio promedio más alto, seguido de cerca por Sarrià-Sant Gervasi. Por otro lado, Nou Barris parece ser el más económico en términos de precio promedio.

# Separación de modelo en sets

Queremos separar nuestro modelo en 3 para poder ver la efectividad de este. Usaremos un set de entrenamiento, otro de test y otro de validación. A medido que avancemos iré comprobando la efectividad del test y una vez veamos que es efectivo haremos la validación global.

Como hemos podido ver en el analisis descriptivo encontrabamos mucha diferencia de valores en *price* dependiendo del barrio en el que estuvieran situadas. Por ello, vamos a hacer una distribución proporcional de valores en cada uno de los sets.

Primer retiramos las variables id para los sets

df <- df |>  
 select(-id,  
 -host\_id)

set.seed(123)  
  
training\_test\_split <- initial\_split(df, prop = 0.8, strata = neighbourhood\_group\_cleansed)  
  
# Crear los data frames de entrenamiento y el resto  
training\_data <- training(training\_test\_split)  
test\_data <- testing(training\_test\_split)  
rm(training\_test\_split)

# Primer modelo

Vamos a hacer una prueba con el primer modelo para ver que resultados obtenemos y como podemos mejorarlo aplicando técnicas estadísticas.

En este caso solo vamos a aplicar las variables numéricas.

training\_data1 <- training\_data %>%   
 select\_if(is.numeric)

Las aplicamos al modelo

modelo1 <- lm(price ~ ., data = training\_data1)

Duplicamos test\_data para esta prueba

test\_data1 <- test\_data

Predecimos con el test\_data:

predicciones\_modelo1 <- predict(modelo1, newdata = test\_data1)

Añadimos las predicciones una columna para comparar

test\_data1 <- test\_data1 %>%  
 mutate(predicciones = predicciones\_modelo1)

Calcular las métricas resultantes la variable price

metrics1 <- test\_data1 %>%  
 metrics(truth = price, estimate = predicciones)

# Asegurarse de que 'predicciones' es un vector numérico sin nombres  
test\_data1$predicciones <- unname(test\_data1$predicciones)

Calculamos el promedio de residuos y lo agregamos al data.frame de metrics

residuo1 <- sum(residuals(modelo1))  
  
residuo1 <- c("residuo", "standard", residuo1)  
  
metrics1 <-rbind(metrics1, residuo1)

# **Transformación de variables en valores numéricos y estandarización**

A continuación, vamos a realizar cambios en las variables que no son de tipo numérico para poder ser incluidas en la regresión lineal:

* Para interactuar con la fecha nos vamos a quedar simplemente con el año de ingreso en la plataforma en host\_since
* Escalamos las variables numéricas
* Convertiremos los booleanos a numéricos
* Usamos datos dummy para las columnas categóricas

training\_data <- training\_data |>  
 mutate(host\_since = as.numeric(format(host\_since, "%Y"))) |>   
 mutate(across(where(is.numeric), scale)) |>   
 mutate(across(where(is.logical), as.numeric))  
  
test\_data <- test\_data |>  
 mutate(host\_since = as.numeric(format(host\_since, "%Y"))) |>   
 mutate(across(where(is.numeric), scale)) |>   
 mutate(across(where(is.logical), as.numeric))  
  
training\_data <- dummy\_cols(training\_data,   
 select\_columns = names(which(sapply(training\_data, is.factor))),  
 remove\_selected\_columns = TRUE)  
  
test\_data <- dummy\_cols(test\_data,   
 select\_columns = names(which(sapply(test\_data, is.factor))),  
 remove\_selected\_columns = TRUE)

# Segundo modelo

# En este caso solo vamos a aplicar las variables numéricas.  
training\_data2 <- training\_data   
  
# Las aplicamos al modelo  
modelo2 <- lm(price ~ ., data = training\_data2)  
  
# Duplicamos test\_data para esta prueba  
test\_data2 <- test\_data  
  
# Predecimos con el test\_data  
predicciones\_modelo2 <- predict(modelo2, newdata = test\_data2)  
  
# Añadimos las predicciones una columna para comparar  
test\_data2 <- test\_data2 %>%  
 mutate(predicciones = predicciones\_modelo2)  
  
# Calcular las métricas resultantes la variable price  
metrics2 <- test\_data2 %>%  
 metrics(truth = price, estimate = predicciones)  
  
# Asegurarse de que 'predicciones' es un vector numérico sin nombres  
test\_data2$predicciones <- unname(test\_data2$predicciones)  
  
# Calculamos el promedio de residuos y lo agregamos al data.frame de metrics  
residuo2 <- sum(residuals(modelo2))  
  
residuo2 <- c("residuo", "standard", residuo2)  
  
metrics2 <-rbind(metrics2, residuo2)

# Análisis de componentes principales

Vamos a realizar el análisis de componentes principales, ya que tenemos un total de 43 variables, entonces vamos a ver cuales son las mas relevantes para nuestro modelo:

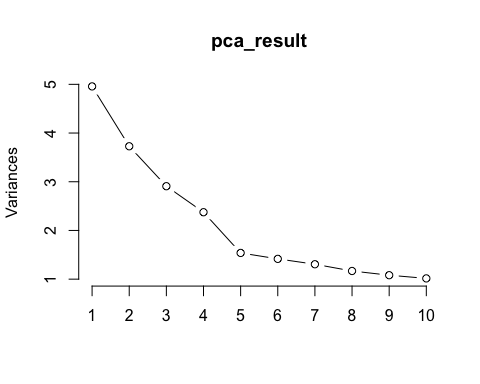
Vamos a utilizar la función pccomp para aplicar el PCA

pca\_result <- prcomp(training\_data, center = TRUE, scale. = FALSE)  
  
summary(pca\_result)

Importance of components:  
 PC1 PC2 PC3 PC4 PC5 PC6 PC7  
Standard deviation 2.2266 1.931 1.70541 1.54084 1.2403 1.19012 1.14289  
Proportion of Variance 0.1569 0.118 0.09207 0.07516 0.0487 0.04484 0.04135  
Cumulative Proportion 0.1569 0.275 0.36704 0.44220 0.4909 0.53574 0.57709  
 PC8 PC9 PC10 PC11 PC12 PC13 PC14  
Standard deviation 1.08057 1.03985 1.00771 0.95473 0.91907 0.87915 0.82879  
Proportion of Variance 0.03696 0.03423 0.03215 0.02886 0.02674 0.02447 0.02174  
Cumulative Proportion 0.61405 0.64828 0.68042 0.70928 0.73602 0.76049 0.78223  
 PC15 PC16 PC17 PC18 PC19 PC20 PC21  
Standard deviation 0.7969 0.77042 0.75043 0.71498 0.70163 0.64249 0.62454  
Proportion of Variance 0.0201 0.01879 0.01783 0.01618 0.01558 0.01307 0.01235  
Cumulative Proportion 0.8023 0.82113 0.83895 0.85513 0.87072 0.88379 0.89613  
 PC22 PC23 PC24 PC25 PC26 PC27 PC28  
Standard deviation 0.61392 0.53499 0.51869 0.50456 0.49769 0.46253 0.44662  
Proportion of Variance 0.01193 0.00906 0.00852 0.00806 0.00784 0.00677 0.00631  
Cumulative Proportion 0.90807 0.91713 0.92564 0.93370 0.94154 0.94832 0.95463  
 PC29 PC30 PC31 PC32 PC33 PC34 PC35  
Standard deviation 0.40400 0.38497 0.36813 0.35794 0.33608 0.33070 0.30716  
Proportion of Variance 0.00517 0.00469 0.00429 0.00406 0.00358 0.00346 0.00299  
Cumulative Proportion 0.95980 0.96449 0.96878 0.97283 0.97641 0.97987 0.98286  
 PC36 PC37 PC38 PC39 PC40 PC41 PC42  
Standard deviation 0.2919 0.27896 0.27066 0.24328 0.23427 0.18280 0.16969  
Proportion of Variance 0.0027 0.00246 0.00232 0.00187 0.00174 0.00106 0.00091  
Cumulative Proportion 0.9856 0.98802 0.99034 0.99221 0.99395 0.99501 0.99592  
 PC43 PC44 PC45 PC46 PC47 PC48 PC49  
Standard deviation 0.15306 0.14298 0.13266 0.12822 0.11115 0.10987 0.10642  
Proportion of Variance 0.00074 0.00065 0.00056 0.00052 0.00039 0.00038 0.00036  
Cumulative Proportion 0.99666 0.99731 0.99786 0.99839 0.99878 0.99916 0.99952  
 PC50 PC51 PC52 PC53 PC54 PC55  
Standard deviation 0.08365 0.07668 0.04854 0.004806 5.147e-15 2.325e-15  
Proportion of Variance 0.00022 0.00019 0.00007 0.000000 0.000e+00 0.000e+00  
Cumulative Proportion 0.99974 0.99992 1.00000 1.000000 1.000e+00 1.000e+00  
 PC56  
Standard deviation 1.759e-15  
Proportion of Variance 0.000e+00  
Cumulative Proportion 1.000e+00

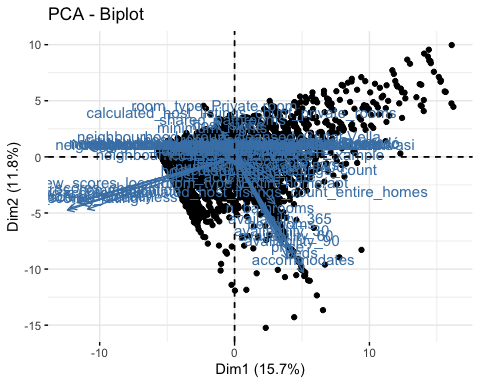
Vamos a hacer un Scree plot para ver cual es el número de componentes principales ideal

scree\_plot <- plot(pca\_result, type = "l")



Hacemos también un biplot para ver las variables correspondientes.

fviz\_pca\_biplot(pca\_result,   
 label="var")



Queremos visualizar las cargas de cada componente y ver que variables tienen mayor influencia.

loadings <- pca\_result$rotation  
  
# Ver las cargas para los primeros cinco componentes principales  
loadings[, 1:5]

PC1 PC2  
host\_since -1.329794e-02 0.0600468772  
host\_response\_rate -1.506824e-02 -0.0827350674  
host\_acceptance\_rate 1.510227e-02 -0.1012475554  
host\_is\_superhost -1.367036e-02 -0.0199229726  
verification\_email 2.763991e-03 -0.0082802230  
verification\_phone 5.304676e-04 -0.0011367906  
verification\_work\_email 1.310210e-02 -0.0246635517  
host\_has\_profile\_pic 1.389175e-04 -0.0015329134  
host\_identity\_verified 1.406812e-02 -0.0359118977  
calculated\_host\_listings\_count 8.885619e-02 -0.0777939849  
calculated\_host\_listings\_count\_entire\_homes 9.014873e-02 -0.1464410138  
calculated\_host\_listings\_count\_private\_rooms 1.582687e-02 0.1000974203  
calculated\_host\_listings\_count\_shared\_rooms 4.646351e-03 0.0003550779  
latitude 2.280411e-03 0.0011033939  
longitude -1.012975e-02 0.0012180506  
accommodates 1.542212e-01 -0.3582400287  
n\_bathrooms 7.952464e-02 -0.1948566849  
shared\_bathrooms -2.628781e-02 0.0783844509  
bedrooms 1.044553e-01 -0.2450313646  
beds 1.498007e-01 -0.3346220856  
price 1.188991e-01 -0.3161530926  
minimum\_nights -4.972958e-02 0.0535750381  
maximum\_nights 2.083702e-02 -0.0232843628  
has\_availability 1.354768e-02 -0.0356863188  
availability\_30 1.058507e-01 -0.2689988576  
availability\_60 1.150830e-01 -0.2823692393  
availability\_90 1.290433e-01 -0.3000048211  
availability\_365 1.021080e-01 -0.2304152371  
instant\_bookable 3.610544e-02 -0.0533977259  
number\_of\_reviews 8.571696e-02 -0.0620207138  
number\_of\_reviews\_l30d 5.699088e-02 -0.0704526095  
review\_scores\_rating -3.742444e-01 -0.1653872979  
review\_scores\_accuracy -3.681967e-01 -0.1590602539  
review\_scores\_cleanliness -3.288750e-01 -0.1621079173  
review\_scores\_checkin -3.415546e-01 -0.1331689109  
review\_scores\_communication -3.525812e-01 -0.1415901566  
review\_scores\_location -3.035424e-01 -0.1188669675  
review\_scores\_value -3.409034e-01 -0.1489737754  
host\_response\_time\_<1h 8.936130e-03 -0.0181814663  
host\_response\_time\_>48h -8.314110e-04 0.0011442877  
host\_response\_time\_~12h -3.948391e-03 0.0033565281  
host\_response\_time\_~24h -4.156328e-03 0.0136806505  
neighbourhood\_group\_cleansed\_Ciutat Vella -1.205371e-02 0.0236200658  
neighbourhood\_group\_cleansed\_Eixample 1.253710e-02 -0.0328179915  
neighbourhood\_group\_cleansed\_Gràcia 1.981800e-03 -0.0009039891  
neighbourhood\_group\_cleansed\_Horta-Guinardó 2.410852e-04 0.0051592680  
neighbourhood\_group\_cleansed\_Les Corts -3.169725e-05 -0.0005420218  
neighbourhood\_group\_cleansed\_Nou Barris -5.695354e-04 0.0023659183  
neighbourhood\_group\_cleansed\_Sant Andreu -1.835902e-03 0.0022127315  
neighbourhood\_group\_cleansed\_Sant Martí -1.560804e-03 -0.0014779770  
neighbourhood\_group\_cleansed\_Sants-Montjuïc 1.185339e-04 0.0038832191  
neighbourhood\_group\_cleansed\_Sarrià-Sant Gervasi 1.173125e-03 -0.0014992234  
room\_type\_Entire home/apt 4.501571e-02 -0.1177308016  
room\_type\_Hotel room -6.299076e-04 -0.0014842007  
room\_type\_Private room -4.505813e-02 0.1197540981  
room\_type\_Shared room 6.723325e-04 -0.0005390958  
 PC3 PC4  
host\_since -0.2088346520 0.0510071397  
host\_response\_rate -0.0539675008 -0.3521861395  
host\_acceptance\_rate -0.0936141586 -0.2398610348  
host\_is\_superhost 0.0136690136 -0.0844453789  
verification\_email 0.0114193918 0.0025701272  
verification\_phone -0.0006214859 -0.0001126510  
verification\_work\_email 0.0321680337 0.0467680109  
host\_has\_profile\_pic 0.0092739799 0.0017763847  
host\_identity\_verified -0.0233973805 -0.0006013192  
calculated\_host\_listings\_count 0.0257247502 0.5280374449  
calculated\_host\_listings\_count\_entire\_homes 0.0811329621 0.4161606737  
calculated\_host\_listings\_count\_private\_rooms -0.0867688093 0.2949745710  
calculated\_host\_listings\_count\_shared\_rooms -0.0282702279 0.0011387480  
latitude 0.0198275337 -0.0376246988  
longitude 0.0090864595 -0.0616367743  
accommodates 0.2895985923 -0.0254703211  
n\_bathrooms 0.1884428122 0.1139661282  
shared\_bathrooms -0.0454644862 -0.0044994108  
bedrooms 0.2760289457 0.0479025241  
beds 0.2816140166 -0.0049099798  
price 0.1599195379 -0.0683033624  
minimum\_nights -0.0480298108 0.1499057048  
maximum\_nights 0.0617814506 -0.0759677650  
has\_availability -0.0231607370 -0.0041115488  
availability\_30 -0.3980835741 0.0024607266  
availability\_60 -0.4338196680 0.0108998670  
availability\_90 -0.4105238328 -0.0063265263  
availability\_365 -0.2657484376 0.0635007750  
instant\_bookable -0.0309782687 -0.0339443865  
number\_of\_reviews 0.1040138507 -0.3353858360  
number\_of\_reviews\_l30d 0.0036729611 -0.2935354540  
review\_scores\_rating 0.0059930924 0.0154329047  
review\_scores\_accuracy 0.0121662861 0.0242194045  
review\_scores\_cleanliness -0.0052088973 0.0245073461  
review\_scores\_checkin -0.0084051610 0.0039899424  
review\_scores\_communication -0.0040230510 0.0092521951  
review\_scores\_location -0.0033231582 0.0589862970  
review\_scores\_value 0.0281256056 -0.0003308220  
host\_response\_time\_<1h 0.0202009453 -0.0472507624  
host\_response\_time\_>48h -0.0033893354 0.0023887155  
host\_response\_time\_~12h -0.0023128998 0.0104544173  
host\_response\_time\_~24h -0.0144987101 0.0344076295  
neighbourhood\_group\_cleansed\_Ciutat Vella -0.0288505034 -0.0008779528  
neighbourhood\_group\_cleansed\_Eixample 0.0392741900 0.0198844444  
neighbourhood\_group\_cleansed\_Gràcia -0.0020103956 -0.0013702546  
neighbourhood\_group\_cleansed\_Horta-Guinardó -0.0019060571 -0.0032291175  
neighbourhood\_group\_cleansed\_Les Corts 0.0013907935 0.0005662008  
neighbourhood\_group\_cleansed\_Nou Barris -0.0028508955 -0.0019151077  
neighbourhood\_group\_cleansed\_Sant Andreu -0.0011854522 -0.0009796646  
neighbourhood\_group\_cleansed\_Sant Martí 0.0025652553 -0.0165772490  
neighbourhood\_group\_cleansed\_Sants-Montjuïc -0.0038138222 -0.0043432135  
neighbourhood\_group\_cleansed\_Sarrià-Sant Gervasi -0.0026131128 0.0088419145  
room\_type\_Entire home/apt 0.0912746021 0.0054439093  
room\_type\_Hotel room -0.0025575390 -0.0028117113  
room\_type\_Private room -0.0859278026 -0.0006926291  
room\_type\_Shared room -0.0027892605 -0.0019395689  
 PC5  
host\_since -0.2204167432  
host\_response\_rate -0.0462318593  
host\_acceptance\_rate 0.1153871668  
host\_is\_superhost 0.0339606899  
verification\_email 0.0114422775  
verification\_phone 0.0010527661  
verification\_work\_email 0.0684347398  
host\_has\_profile\_pic 0.0095338017  
host\_identity\_verified 0.0279872719  
calculated\_host\_listings\_count 0.3282649777  
calculated\_host\_listings\_count\_entire\_homes 0.4453477339  
calculated\_host\_listings\_count\_private\_rooms -0.1293635123  
calculated\_host\_listings\_count\_shared\_rooms -0.0622836078  
latitude -0.0788645990  
longitude -0.0362393383  
accommodates -0.0861829186  
n\_bathrooms -0.3351919991  
shared\_bathrooms -0.0721513994  
bedrooms -0.3182836001  
beds -0.1534420560  
price 0.0297888200  
minimum\_nights -0.1667698788  
maximum\_nights 0.0027743017  
has\_availability 0.0270989318  
availability\_30 -0.1067672933  
availability\_60 -0.0565450039  
availability\_90 0.0094036250  
availability\_365 0.0586712414  
instant\_bookable 0.0693660438  
number\_of\_reviews 0.3606523420  
number\_of\_reviews\_l30d 0.3560720235  
review\_scores\_rating 0.0374610612  
review\_scores\_accuracy 0.0336798345  
review\_scores\_cleanliness 0.0495934344  
review\_scores\_checkin -0.0221275293  
review\_scores\_communication 0.0088569349  
review\_scores\_location -0.0187894211  
review\_scores\_value 0.0374091340  
host\_response\_time\_<1h 0.0545077367  
host\_response\_time\_>48h -0.0070998954  
host\_response\_time\_~12h -0.0202826472  
host\_response\_time\_~24h -0.0271251942  
neighbourhood\_group\_cleansed\_Ciutat Vella -0.0179801694  
neighbourhood\_group\_cleansed\_Eixample 0.0345009163  
neighbourhood\_group\_cleansed\_Gràcia 0.0081219490  
neighbourhood\_group\_cleansed\_Horta-Guinardó -0.0063807623  
neighbourhood\_group\_cleansed\_Les Corts -0.0009901357  
neighbourhood\_group\_cleansed\_Nou Barris -0.0054226138  
neighbourhood\_group\_cleansed\_Sant Andreu -0.0048427055  
neighbourhood\_group\_cleansed\_Sant Martí -0.0128668951  
neighbourhood\_group\_cleansed\_Sants-Montjuïc 0.0137072066  
neighbourhood\_group\_cleansed\_Sarrià-Sant Gervasi -0.0078467900  
room\_type\_Entire home/apt 0.1062638271  
room\_type\_Hotel room 0.0016725317  
room\_type\_Private room -0.1045304097  
room\_type\_Shared room -0.0034059491

# Convertir la matriz de cargas en un data frame  
loadings\_df <- as.data.frame(loadings) |>  
 select(PC1, PC2, PC3, PC4, PC5)  
  
# Sumamos los PC de cada variable  
loadings\_df$Suma <- rowSums(loadings\_df[, c("PC1", "PC2", "PC3", "PC4", "PC5")])   
  
loadings\_df\_sorted <- loadings\_df |>  
 arrange(desc(Suma))

El resultado es bastante sorprendente, pero tiene sentido vemos que las variables mas relevantes en primer lugar son el número de propiedades que tiene el host en si. Esto se puede deber a que a mayor número de propiedades hablamos de una dedicación profesional y conlleva un aumento de precios en sus alojamientos en general. Otras variables como el número de reviews influencia también.

Vamos a crear una lista con las variables mas relevantes, para el siguiente modelo probaremos con las variables cuya suma de componentes sea mayor que 0:

loadings\_df <- loadings\_df\_sorted |>  
 filter(Suma > 0)   
  
seleccion\_variables <- rownames(loadings\_df)  
seleccion\_variables <- c(seleccion\_variables, "price")

# Tercer modelo

Vamos a hacer un modelo solo con las variables seleccionadas:

# En este caso solo vamos a aplicar las variables numéricas.  
training\_data3 <- training\_data |>  
 select(seleccion\_variables)

Warning: Using an external vector in selections was deprecated in tidyselect 1.1.0.  
ℹ Please use `all\_of()` or `any\_of()` instead.  
 # Was:  
 data %>% select(seleccion\_variables)  
  
 # Now:  
 data %>% select(all\_of(seleccion\_variables))  
  
See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.

# Las aplicamos al modelo  
modelo3 <- lm(price ~ ., data = training\_data3)  
  
# Duplicamos test\_data para esta prueba  
test\_data3 <- test\_data |>  
 select(seleccion\_variables)  
  
# Predecimos con el test\_data  
predicciones\_modelo3 <- predict(modelo3, newdata = test\_data3)  
  
# Añadimos las predicciones una columna para comparar  
test\_data3 <- test\_data3 %>%  
 mutate(predicciones = predicciones\_modelo2)  
  
# Calcular las métricas resultantes la variable price  
metrics3 <- test\_data3 %>%  
 metrics(truth = price, estimate = predicciones)  
  
# Asegurarse de que 'predicciones' es un vector numérico sin nombres  
test\_data3$predicciones <- unname(test\_data3$predicciones)  
  
# Calculamos el promedio de residuos y lo agregamos al data.frame de metrics  
residuo3 <- sum(residuals(modelo3))  
  
residuo3 <- c("residuo", "standard", residuo3)  
  
metrics3 <-rbind(metrics3, residuo3)

# Cuarto y quinto modelo

Vamos a hacer dos modelos mas: Uno con una selección mas amplia de variables y otra con una selección mas corta, al final de todo comparemos todos los modelos y sacaremos conclusiones.

## Cuarto modelo

Vamos a hacer una selección mas corta de variables filtrando por 0.10

loadings\_df <- loadings\_df\_sorted |>  
 filter(Suma > 0.10)   
  
seleccion\_variables <- rownames(loadings\_df)  
seleccion\_variables <- c(seleccion\_variables, "price")

# En este caso solo vamos a aplicar las variables numéricas.  
training\_data4 <- training\_data |>  
 select(seleccion\_variables)  
  
# Las aplicamos al modelo  
modelo4 <- lm(price ~ ., data = training\_data4)  
  
# Duplicamos test\_data para esta prueba  
test\_data4 <- test\_data |>  
 select(seleccion\_variables)  
  
# Predecimos con el test\_data  
predicciones\_modelo4 <- predict(modelo4, newdata = test\_data4)  
  
# Añadimos las predicciones una columna para comparar  
test\_data4 <- test\_data4 %>%  
 mutate(predicciones = predicciones\_modelo4)  
  
# Calcular las métricas resultantes la variable price  
metrics4 <- test\_data4 %>%  
 metrics(truth = price, estimate = predicciones)  
  
  
# Asegurarse de que 'predicciones' es un vector numérico sin nombres  
test\_data4$predicciones <- unname(test\_data4$predicciones)  
  
# Calculamos el promedio de residuos y lo agregamos al data.frame de metrics  
residuo4 <- sum(residuals(modelo4))  
  
residuo4 <- c("residuo", "standard", residuo4)  
  
metrics4 <-rbind(metrics4, residuo4)

## Quinto modelo

Ahora haremos una selección mas larga filtrando por -0.10

loadings\_df <- loadings\_df\_sorted |>  
 filter(Suma > -0.20)   
  
seleccion\_variables <- rownames(loadings\_df)  
seleccion\_variables <- c(seleccion\_variables, "price")

training\_data5 <- training\_data |>  
 select(seleccion\_variables)  
  
# Las aplicamos al modelo  
modelo5 <- lm(price ~ ., data = training\_data5)  
  
# Duplicamos test\_data para esta prueba  
test\_data5 <- test\_data |>  
 select(seleccion\_variables)  
  
# Predecimos con el test\_data  
predicciones\_modelo5 <- predict(modelo5, newdata = test\_data5)  
  
# Añadimos las predicciones una columna para comparar  
test\_data5 <- test\_data5 %>%  
 mutate(predicciones = predicciones\_modelo5)  
  
# Calcular las métricas resultantes la variable price  
metrics5 <- test\_data5 %>%  
 metrics(truth = price, estimate = predicciones)  
  
# Asegurarse de que 'predicciones' es un vector numérico sin nombres  
test\_data5$predicciones <- unname(test\_data5$predicciones)  
  
# Calculamos el promedio de residuos y lo agregamos al data.frame de metrics  
residuo5 <- sum(residuals(modelo5))  
  
residuo5 <- c("residuo", "standard", residuo5)  
  
metrics5 <-rbind(metrics5, residuo5)

# Comparativa de modelos

Unficamos las métricas de los 5 modelos:

metricas <- bind\_rows(metrics1, metrics2, metrics3, metrics4, metrics5) |>  
 select(-.estimator)  
n\_modelo <- rep(seq(5), each = 4)  
metricas$n\_modelo <- n\_modelo  
  
metricas <- metricas %>%  
 pivot\_wider(names\_from = .metric, values\_from = .estimate)  
  
print(metricas)

# A tibble: 5 × 5  
 n\_modelo rmse rsq mae residuo   
 <int> <chr> <chr> <chr> <chr>   
1 1 92.2370396724672 0.395825071402194 55.1237606977542 -8.12008238426…  
2 2 0.750713817706505 0.43642217949015 0.444070417275347 -1.05693231944…  
3 3 0.750713817706505 0.43642217949015 0.444070417275347 -1.49963375051…  
4 4 0.898035879986293 0.193568100662718 0.56628162831109 5.153100168797…  
5 5 0.763072846484759 0.417753707452463 0.455242020155687 -4.17055279200…

Vamos a comentar cada modelo basándonos en las métricas proporcionadas y luego seleccionaremos el mejor modelo.

**Modelo 1**

* El RMSE es muy alto, lo que indica que los errores de predicción son, en promedio, grandes.
* Rsq es razonablemente alto, lo que sugiere que el modelo explica una buena cantidad de la varianza de los datos.
* MAE es bastante alto, lo que indica errores significativos en la predicción.
* El residuo es negativo y muy cercano a cero, lo que sugiere que en promedio, el modelo podría estar subestimando ligeramente las predicciones.

**Modelo 2**

* El RMSE es considerablemente más bajo que el del Modelo 1, lo que indica una reducción significativa en los errores de predicción en comparación con el modelo que no utiliza normalización ni variables dummy.
* El Rsq, es menor que el del Modelo 1. Un valor más bajo aquí sugiere que, aunque las predicciones son más precisas en términos de error cuadrático, la proporción de la varianza total que el modelo puede explicar ha disminuido.
* El MAE reducido en el Modelo 2 refleja la mejora en la precisión de las predicciones después de la normalización y la inclusión de variables dummy. Como el MAE no da tanto peso a los errores más grandes como el

**Modelo 3**

* El RMSE y el MAE son idénticos a los del Modelo 2, lo que indica que la selección de componentes no perjudicó la precisión de la predicción.
* El Rsq permanece sin cambios respecto al Modelo 2, lo que sugiere que la cantidad de varianza explicada por el modelo es similar.
* El residuo es negativo y muy pequeño, lo cual es consistente con el Modelo 2.

**Modelo 4**

* Este modelo muestra un RMSE ligeramente más alto y un Rsq ligeramente más bajo que los Modelos 2 y 3, lo que indica que la eliminación de algunas variables mediante un umbral de carga más alto ha tenido un pequeño impacto negativo en el rendimiento del modelo.
* El MAE es ligeramente más alto, lo que sugiere que las predicciones son menos precisas en promedio.
* El residuo es positivo, lo que indica que este modelo podría estar sobreestimando ligeramente las predicciones en promedio.

**Modelo 5**

* Este modelo tiene la menor RMSE y MAE de todos los modelos, lo que sugiere que tiene el mejor rendimiento en términos de precisión de la predicción.
* El R² es también el más alto, indicando que este modelo explica la mayor parte de la varianza en los datos.
* El residuo es negativo, pero muy cercano a cero, similar a los Modelos 2 y 3.

## Selección del modelo

El Modelo 5 parece ser el mejor en términos de todas las métricas proporcionadas. Tiene el RMSE más bajo, lo que indica que tiene el menor error de predicción promedio. También tiene el MAE más bajo, lo que sugiere que es consistente en su precisión a través de diferentes muestras. Además, el R² más alto indica que explica la mayor cantidad de varianza en la variable dependiente comparado con los otros modelos. Aunque el residuo es negativo, lo cual puede implicar una leve subestimación, está muy cerca de cero, lo que sugiere que en promedio, el sesgo del modelo es mínimo.

La inclusión de más variables (con cargas mayores de -0.10) en el Modelo 5 parece haber capturado mejor la complejidad subyacente de los datos, lo que ha resultado en un modelo más preciso.

Sin embargo, en la práctica, se recomendaría realizar una validación cruzada o pruebas en un conjunto de datos de prueba separado para evaluar mejor la capacidad del modelo para generalizar a nuevos datos antes de finalizar la selección del modelo.