Trabajo Final

2024-05-23

```r
#install.packages("codpostal")
library(here)
```

```
## here() starts at C:/Users/gmiranda/OneDrive - Telefonica/Desktop/Facultad/Trabajo_Final_modelos/TRABA
```

```r
library(readxl)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```r
library(ggplot2)
library(visdat)
library(patchwork)
```

```r
df = read_excel(here("airbnb_barcelona_v2.xlsx"))
```

```
## Warning: Expecting numeric in D10008 / R10008C4: got '08001'
```

```
## Warning: Expecting numeric in D11270 / R11270C4: got 'barcelona'
```

```
## Warning: Expecting numeric in D11553 / R11553C4: got '13-08008'
```

```
## Warning: Expecting numeric in D11554 / R11554C4: got '13-08008'
```

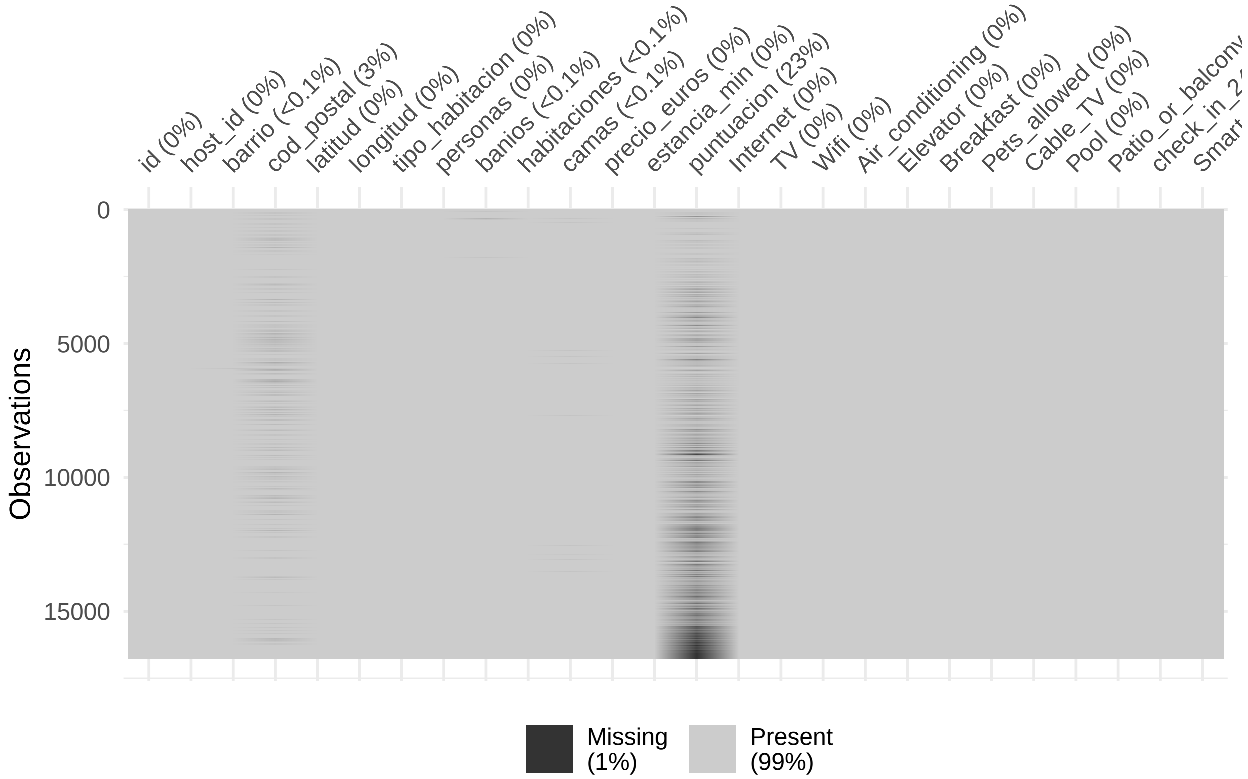
```r
head(df)
```

```
## # A tibble: 6 x 26
##      id host_id barrio      cod_postal latitud longitud tipo_habitacion personas
##   <dbl>   <dbl> <chr>            <dbl>   <dbl>    <dbl> <chr>              <dbl>
## 1 18666   71615 Sant Marta        8026    41.4     2.19 Entire home/apt        6
## 2 18674   71615 La Sagrada~       8025    41.4     2.17 Entire home/apt        8
## 3 21605   82522 Sant Marta        8018    41.4     2.20 Private room           2
## 4 23197   90417 Sant Marta        8930    41.4     2.22 Entire home/apt        6
## 5 25786  108310 Vila de Gr~       8012    41.4     2.16 Private room           2
## 6 31377  134698 Horta-Guin~       8025    41.4     2.17 Private room           2
## # i 18 more variables: banios <dbl>, habitaciones <dbl>, camas <dbl>,
## #   precio_euros <dbl>, estancia_min <dbl>, puntuacion <dbl>, Internet <dbl>,
## #   TV <dbl>, Wifi <dbl>, Air_conditioning <dbl>, Elevator <dbl>,
## #   Breakfast <dbl>, Pets_allowed <dbl>, Cable_TV <dbl>, Pool <dbl>,
## #   Patio_or_balcony <dbl>, check_in_24_hs <dbl>, Smart_lock <dbl>
```

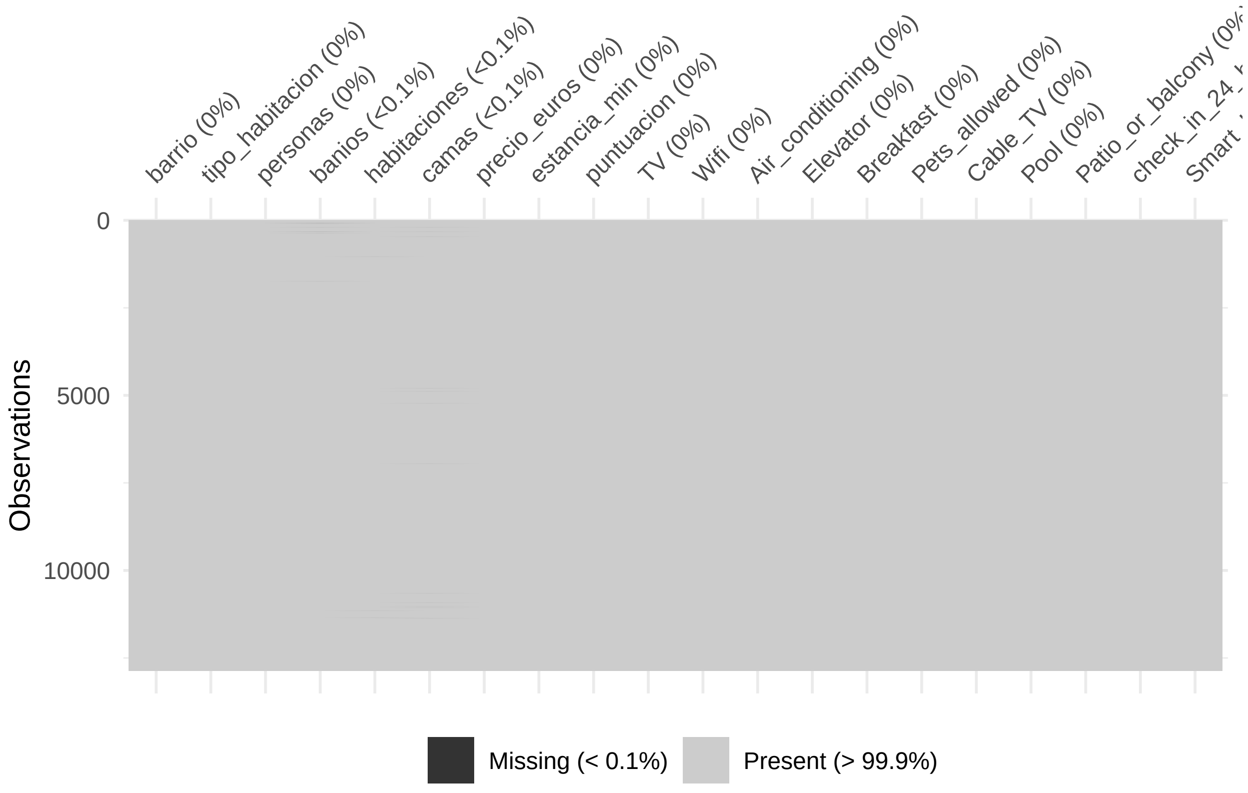
```
summary(df)
```

```
##        id              host_id             barrio            cod_postal
##  Min.   :   18666   Min.   :    10704   Length:16761       Min.   :      0
##  1st Qu.:11448633   1st Qu.:  7612142   Class :character   1st Qu.:   8004
##  Median :22146039   Median : 45072553   Mode  :character   Median :   8012
##  Mean   :20880757   Mean   : 86673374                      Mean   :   8267
##  3rd Qu.:31623085   3rd Qu.:158838753                      3rd Qu.:   8022
##  Max.   :36582760   Max.   :274862556                      Max.   :4008009
##                                                            NA’s   :506
##     latitud         longitud     tipo_habitacion       personas
##  Min.   :41.35   Min.   :2.105   Length:16761       Min.   : 1.000
##  1st Qu.:41.38   1st Qu.:2.157   Class :character   1st Qu.: 2.000
##  Median :41.39   Median :2.168   Mode  :character   Median : 2.000
##  Mean   :41.39   Mean   :2.168                      Mean   : 3.358
##  3rd Qu.:41.40   3rd Qu.:2.178                      3rd Qu.: 4.000
##  Max.   :41.46   Max.   :2.222                      Max.   :18.000
##
##      banios       habitaciones        camas         precio_euros
##  Min.   :0.000   Min.   : 0.000   Min.   : 0.000   Min.   :   7
##  1st Qu.:1.000   1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.:  40
##  Median :1.000   Median : 1.000   Median : 2.000   Median :  63
##  Mean   :1.288   Mean   : 1.586   Mean   : 2.239   Mean   :  92
##  3rd Qu.:1.500   3rd Qu.: 2.000   3rd Qu.: 3.000   3rd Qu.: 107
##  Max.   :8.000   Max.   :12.000   Max.   :30.000   Max.   :1000
##  NA’s   :9       NA’s   :3        NA’s   :16
##   estancia_min       puntuacion        Internet           TV
##  Min.   :  1.000   Min.   : 20.00   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:  1.000   1st Qu.: 88.00   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :  2.000   Median : 93.00   Median :0.0000   Median :1.0000
##  Mean   :  8.509   Mean   : 90.98   Mean   :0.2149   Mean   :0.6973
##  3rd Qu.:  4.000   3rd Qu.: 97.00   3rd Qu.:0.0000   3rd Qu.:1.0000
##  Max.   :900.000   Max.   :100.00   Max.   :1.0000   Max.   :1.0000
##                    NA’s   :3891
##       Wifi        Air_conditioning    Elevator        Breakfast
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000
##  Median :1.0000   Median :1.0000   Median :1.0000   Median :0.00000
##  Mean   :0.7383   Mean   :0.5707   Mean   :0.6167   Mean   :0.05913
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.00000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.00000
##
##   Pets_allowed       Cable_TV            Pool         Patio_or_balcony
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000
##  Median :0.0000   Median :0.00000   Median :0.00000   Median :0.0000
##  Mean   :0.1157   Mean   :0.09898   Mean   :0.01873   Mean   :0.2261
##  3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.00000   Max.   :1.0000
##
##  check_in_24_hs     Smart_lock
##  Min.   :0.0000   Min.   :0.000000
##  1st Qu.:0.0000   1st Qu.:0.000000
##  Median :0.0000   Median :0.000000
##  Mean   :0.1107   Mean   :0.007458
##  3rd Qu.:0.0000   3rd Qu.:0.000000
##  Max.   :1.0000   Max.   :1.000000
##
```

```
df%>%vis_miss()
```



```
df= df %>% select(barrio,
                  tipo_habitacion,
                  personas,
                  banios,
                  habitaciones,
                  camas,
                  precio_euros,
                  estancia_min,
                  puntuacion,
                  TV,
                  Wifi,
                  Air_conditioning,
                  Elevator,
                  Breakfast,
                  Pets_allowed,
                  Cable_TV,
                  Pool,
                  Patio_or_balcony,
                  check_in_24_hs,
                  Smart_lock)%>% filter(is.na(puntuacion)==FALSE,
                                        is.na(barrio)==FALSE)

df%>%vis_miss()
```



```
#df%>% summarise(sum_wifi=sum(Wifi),sum_int=sum(Internet))
```

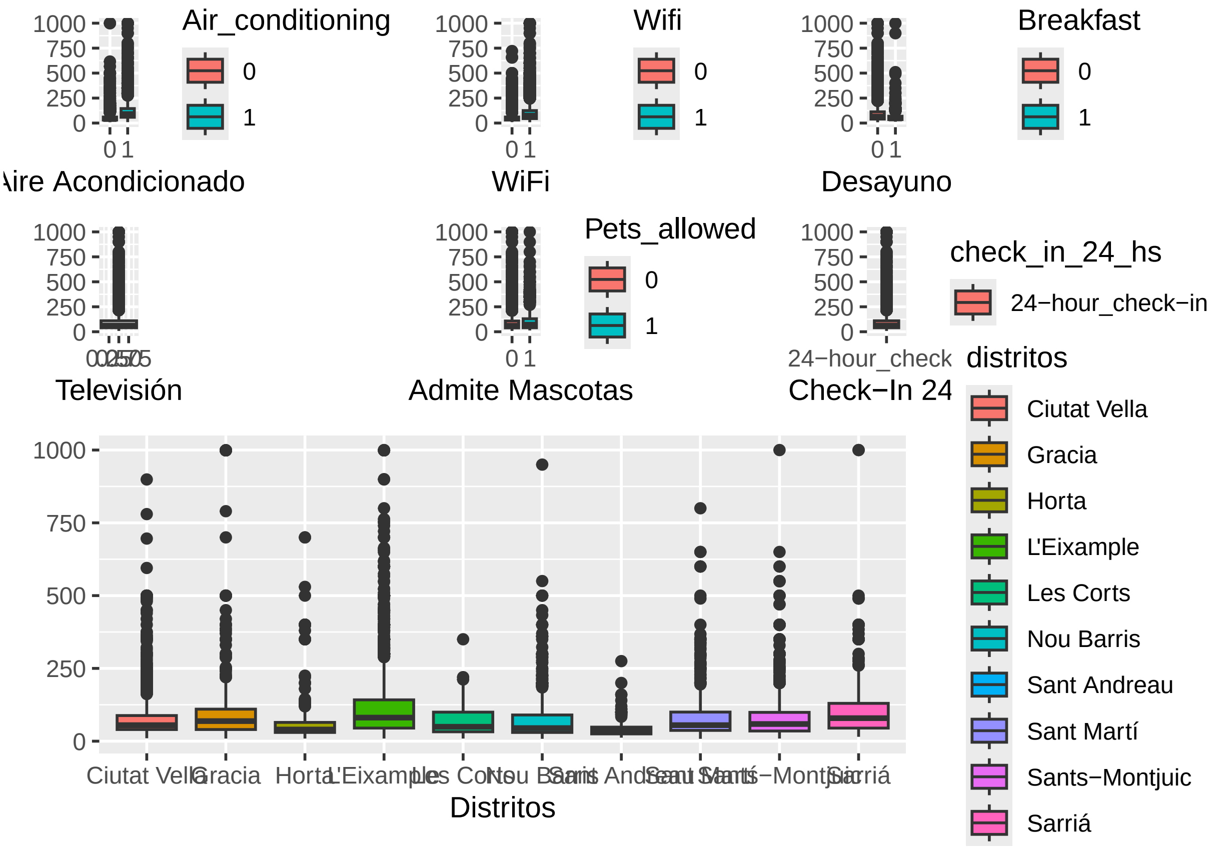
```
distritos<-character(length(nrow(df)))


#CON AYUDA, primero hay que sacar las observaciones que no tienen el dato del barrio para que funcione

for (i in seq_along(df$barrio)) {
```

```r
    if (df$barrio[i] == "Sant Marta"||
               df$barrio[i] == "La Guineueta - Canyelles"||
               df$barrio[i] == "La Prosperitat"||
               df$barrio[i] == "Nou Barris"||
               df$barrio[i] == "Porta"||
               df$barrio[i] == "Trinitat Nova"||
               df$barrio[i] == "Turo de la Peira - Can Peguera"||
               df$barrio[i] == "Verdum - Los Roquetes"||
               df$barrio[i] == "Vilapicina i la Torre Llobeta") {
      distritos[i] <- "Nou Barris"
    } else if (df$barrio[i]=='La Sagrada Familia' ||
               df$barrio[i] == "Eixample" ||
               df$barrio[i] == "L'Antiga Esquerra de l'Eixample" ||
               df$barrio[i] == "Sant Antoni" ||
               df$barrio[i] == "Dreta de l'Eixample" ||
               df$barrio[i] == "La Nova Esquerra de l'Eixample" ||
               df$barrio[i] == "el Fort Pienc") {
      distritos[i] <- "L'Eixample"
    } else if (df$barrio[i] == "Vila de Gracia" ||
               df$barrio[i] == "Camp d'en Grassot i Gracia Nova" ||
               df$barrio[i] == "Gracia" ||
               df$barrio[i] == "El Coll" ||
               df$barrio[i] == "La Salut" ||
               df$barrio[i] == "Vallcarca i els Penitents") {
      distritos[i] <- "Gracia"
    } else if (df$barrio[i] == "Horta-Guinarda" ||
               df$barrio[i] == "Can Baro" ||
               df$barrio[i] == "Carmel" ||
               df$barrio[i] == "El Baix Guinardo"||
               df$barrio[i] == "Guinarda" ||
               df$barrio[i] == "Horta"||
               df$barrio[i] == "La Font d'en Fargues"||
               df$barrio[i] == "La Teixonera"||
               df$barrio[i] == "La Vall d'Hebron"||
               df$barrio[i] == "Montbau"||
               df$barrio[i] == "Sant Genis dels Agudells") {
      distritos[i] <- "Horta"
    } else if (df$barrio[i] == "Les Corts"||
               df$barrio[i] == "La Maternitat i Sant Ramon"||
               df$barrio[i] == "Pedralbes") {
      distritos[i] <- "Les Corts"
    } else if (df$barrio[i] == "El Gotic" ||
               df$barrio[i] == "La Barceloneta" ||
               df$barrio[i] == "Ciutat Vella" ||
               df$barrio[i] == "El Raval" ||
               df$barrio[i] == "Sant Pere/Santa Caterina" ||
               df$barrio[i] == "El Born") {
      distritos[i] <- "Ciutat Vella"
    } else if (df$barrio[i] == "El Poble-sec" ||
               df$barrio[i] == "Sants-Montjuic") {
      distritos[i] <- "Sants-Montjuic"
    } else if (df$barrio[i] == "El Clot" ||
               df$barrio[i] == "El Besos i el Maresme" ||
               df$barrio[i] == "El Camp de l'Arpa del Clot"||
               df$barrio[i] == "El Poblenou"||
               df$barrio[i] == "La Vila Olimpica"||
               df$barrio[i] == "Diagonal Mar - La Mar Bella"||
               df$barrio[i] == "Glaries - El Parc"||
               df$barrio[i] == "La Verneda i La Pau"||
               df$barrio[i] == "Provencals del Poblenou"||
               df$barrio[i] == "Sant Marta de Provencals"
               ) {
      distritos[i] <- "Sant Martí"
    } else if (df$barrio[i] == "Sant Gervasi - Galvany"||
               df$barrio[i] == "El Putget i Farro"||
               df$barrio[i] == "Les Tres Torres"||
               df$barrio[i] == "Sant Gervasi - la Bonanova"||
               df$barrio[i] == "Sarria"||
               df$barrio[i] == "Sarria-Sant Gervasi") {
      distritos[i] <- "Sarriá"
    } else if (df$barrio[i] == "El Bon Pastor" ||
               df$barrio[i] == "El Congres i els Indians"||
               df$barrio[i] == "La Sagrera"||
               df$barrio[i] == "La Trinitat Vella"||
               df$barrio[i] == "Navas"||
               df$barrio[i] == "Sant Andreu"||
               df$barrio[i] == "Sant Andreu de Palomar") {
      distritos[i] <- "Sant Andreau"
  }
}
df$distritos <- distritos
```

```
summary(df)
```

```
##     barrio          tipo_habitacion       personas         banios
##  Length:12869       Length:12869       Min.   : 1.00   Min.   :0.000
##  Class :character   Class :character   1st Qu.: 2.00   1st Qu.:1.000
##  Mode  :character   Mode  :character   Median : 2.00   Median :1.000
##                                        Mean   : 3.41   Mean   :1.281
##                                        3rd Qu.: 4.00   3rd Qu.:1.500
##                                        Max.   :18.00   Max.   :7.500
##                                                        NA’s   :8
##   habitaciones        camas         precio_euros      estancia_min
##  Min.   : 0.000   Min.   : 0.000   Min.   :   8.00   Min.   :  1.000
##  1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.:  40.00   1st Qu.:  1.000
##  Median : 1.000   Median : 2.000   Median :  62.00   Median :  2.000
##  Mean   : 1.595   Mean   : 2.261   Mean   :  92.44   Mean   :  5.973
##  3rd Qu.: 2.000   3rd Qu.: 3.000   3rd Qu.: 110.00   3rd Qu.:  3.000
##  Max.   :12.000   Max.   :26.000   Max.   :1000.00   Max.   :365.000
##  NA’s   :3        NA’s   :12
##    puntuacion           TV              Wifi        Air_conditioning
##  Min.   : 20.00   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.: 88.00   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median : 93.00   Median :1.0000   Median :1.0000   Median :1.0000
##  Mean   : 90.98   Mean   :0.6954   Mean   :0.7467   Mean   :0.5722
##  3rd Qu.: 97.00   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :100.00   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##
##     Elevator        Breakfast        Pets_allowed       Cable_TV
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :1.0000   Median :0.00000   Median :0.0000   Median :0.0000
##  Mean   :0.6264   Mean   :0.05719   Mean   :0.1111   Mean   :0.1133
##  3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000
##
##       Pool          Patio_or_balcony check_in_24_hs     Smart_lock
##  Min.   :0.00000   Min.   :0.00     Min.   :0.0000   Min.   :0.000000
##  1st Qu.:0.00000   1st Qu.:0.00     1st Qu.:0.0000   1st Qu.:0.000000
##  Median :0.00000   Median :0.00     Median :0.0000   Median :0.000000
##  Mean   :0.01702   Mean   :0.25     Mean   :0.1225   Mean   :0.008703
##  3rd Qu.:0.00000   3rd Qu.:0.00     3rd Qu.:0.0000   3rd Qu.:0.000000
##  Max.   :1.00000   Max.   :1.00     Max.   :1.0000   Max.   :1.000000
##
##   distritos
##  Length:12869
##  Class :character
##  Mode  :character
##
##
##
##
```

```
df_final= df%>% mutate(tipo_habitacion=as.factor(tipo_habitacion),
                       banios=as.factor(banios),
                       camas=as.factor(camas),
                       habitaciones=as.factor(habitaciones),
                       distritos=as.factor(distritos),
                       wifi=as.factor(Wifi),
                      tv=as.factor(TV),
                  Wifi=as.factor(wifi),
                  Air_conditioning=as.factor(Air_conditioning),
                  Elevator=as.factor(Elevator),
                  Breakfast=as.factor(Breakfast),
                  Pets_allowed=as.factor(Pets_allowed),
                  Cable_TV=as.factor(Cable_TV),
                  Pool=as.factor(Pool),
                  Patio_or_balcony=as.factor(Patio_or_balcony),
                  check_in_24_hs=as.factor("24-hour_check-in"),
                  Smart_lock=as.factor(Smart_lock)
             )
```



```
  # Agregar el vector distritos al data.frame


#miramos si hay valores de las variables explicativas para todos los id

airbnb_barcelona %>% filter(is.na(puntuacion)) %>% select(puntuacion) #3891 obs sin valor
airbnb_barcelona %>% filter(is.na(barrio)) %>% select(barrio) #1 sin valor
airbnb_barcelona %>% filter(is.na(personas)) %>% select(personas) #info de todas
airbnb_barcelona %>% filter(is.na(banios)) %>% select(banios) #9 sin valor
airbnb_barcelona %>% filter(is.na(habitaciones)) %>% select(habitaciones) #3 sin valor
airbnb_barcelona %>% filter(is.na(camas)) %>% select(camas) #16 sin valor
airbnb_barcelona %>% filter(is.na(precio_euros)) %>% select(precio_euros) #info de todas
airbnb_barcelona %>% filter(is.na(estancia_min)) %>% select(estancia_min) #info de todas
airbnb_barcelona %>% filter(is.na(amenities)) %>% select(amenities) #info de todas
airbnb_barcelona %>% filter(is.na(tipo_habitacion)) %>% select(tipo_habitacion) #info de todas
airbnb_barcelona %>% filter(is.na(longitud)) %>% select(longitud) #info de todas
airbnb_barcelona %>% filter(is.na(latitud)) %>% select(latitud) #info de todas
airbnb_barcelona %>% filter(is.na(cod_postal)) %>% select(cod_postal) #506 sin valor


#SELECCIÓN DE VARIABLES

k=ncol(airbnb_barcelona)-1
modelos_posibles=2**k-1

#hay 16383 modelos posibles, vamos a aplicar los procedimientos de hipotesis para
```

```
#llegar al mejor modelo
library(tidyverse)

#SI HAGO mod0 SE ROMPE R)??
mod0 <- lm(precio_euros ~  distritos + tipo_habitacion + personas + banios + habitaciones + camas + amen
coef(mod0)
summary(mod0)

#CUAL ES LA LIBRERIA??
#forward
modF <- forward(mod0, alpha = 0.05)
length(coef(modF))  # parametros
summary(modF)

# backward
modB <- backward(mod0, alpha = 0.01)
length(coef(modB))  # parametros
summary(modB)

# stepwise
modS <- stepWise(mod0, alpha.enter = 0.04, alpha.remove=0.05)
length(coef(modS))  # parametros
summary(modS)

#COMPARAMOS POR AIC, BIC O R2 AJUSTADO PARA VER CUAL ES EL MEJOR

#que paquete se usaba para ver la tabla con AIC BIC
install.packages("HH")
library(HH)
summaryHH(modF)
summaryHH(modB)

help(anova)
summaryHH(modS)
```