# Pràctica 2: Neteja i anàlisi de les dades

Francesc Ballester Lecina - Oriol Raurell Gan

## Contents



# Introducció, descripció i objectius de l'activitat

## Introducció

La problemàtica de l'augment del nombre d'allotjaments turístics en grans ciutats com ara Barcelona ha provocat canvis en la legislació per tal de regular el sector. Estudiar una mostra de diferents allotjaments que s'ofereixen en les plataformes més importants, com per exemple Airbnb, ens permet entendre quines són les característiques que disposa l'allotjament turístic més habitual així com les variables que determinen el preu final per nit.

## Descripció i objectius

El dataset original s'ha obtingut de la plataforma kaggle:

https://www.kaggle.com/datamarket/alojamientos-tursticos

A continuació es mostra el dataset original al complet i el resum de les variables:

```
# Importem el joc de dades original
data<-read.csv("./alojamientos-turisticos-sample.csv", sep=",", header=T)

# Resum de les variables
str(data)
```

```
## 'data.frame':    10000 obs. of  40 variables:
##  $ apartment_id               : int  26316169 34672572 32603220 3181996 5914200 11355385 14703586 169
##  $ url                        : chr  "https://www.airbnb.com/rooms/26316169" "https://www.airbnb.com/
##  $ name                       : chr  "Olivia's Place" "1Bedroom apartment in Atocha - Wanda Metropoli
##  $ description                : chr  "Olivia‘s place es Piso reformado a nuevo pensando en huÃ©spedes
##  $ host_id                    : int  123008256 3256859 107752533 16133934 11857176 24995622 81612614
##  $ neighborhood_overview      : chr  "Barrio tranquilo, hay un chino al lado que vende latas de cerve
##  $ neighbourhood_name         : chr  "Este" "Cortes" "Felanitx" "el Barri GÃ²tic" ...
##  $ neighbourhood_district     : chr  "" "Centro" "" "Ciutat Vella" ...
##  $ latitude                   : num  36.7 40.4 39.4 41.4 41.8 ...
##  $ longitude                  : num  -4.36 -3.7 3.24 2.18 3.07 ...
##  $ room_type                  : chr  "Entire home/apt" "Entire home/apt" "Entire home/apt" "Private r
##  $ accommodates               : int  5 4 2 2 3 2 6 6 5 7 ...
##  $ bathrooms                  : int  1 1 1 1 1 1 2 1 1 2 ...
##  $ bedrooms                   : int  3 1 1 1 1 1 3 3 NA 3 ...
##  $ beds                       : int  4 2 2 1 2 1 5 4 3 6 ...
##  $ amenities_list             : chr  "{TV,Internet,Wifi,\"Air conditioning\",Kitchen,Breakfast,\"Free
##  $ price                      : num  80 500 60 48 99 25 103 100 33 99 ...
##  $ minimum_nights             : int  2 1 2 1 6 2 1 4 3 2 ...
##  $ maximum_nights             : int  30 1125 15 12 31 1125 1125 28 380 60 ...
##  $ has_availability           : chr  "true" "true" "true" "true" ...
##  $ availability_30            : int  0 0 30 5 30 7 29 25 1 7 ...
##  $ availability_60            : int  0 0 59 35 60 23 59 55 1 18 ...
##  $ availability_90            : int  0 0 59 65 90 43 89 85 1 27 ...
##  $ availability_365           : int  167 0 59 155 181 43 233 330 1 62 ...
##  $ number_of_reviews          : int  11 0 15 173 0 221 3 47 0 4 ...
##  $ first_review_date          : chr  "2018-07-06" "" "2019-04-07" "2014-08-11" ...
##  $ last_review_date           : chr  "2019-08-25" "" "2019-09-09" "2018-09-25" ...
##  $ review_scores_rating       : num  96 NA 96 91 NA 96 73 100 NA 85 ...
##  $ review_scores_accuracy     : num  10 NA 10 9 NA 10 8 10 NA 9 ...
##  $ review_scores_cleanliness  : num  10 NA 9 10 NA 10 7 10 NA 10 ...
##  $ review_scores_checkin      : num  10 NA 10 9 NA 10 9 10 NA 10 ...
##  $ review_scores_communication: num  10 NA 10 9 NA 10 7 10 NA 9 ...
##  $ review_scores_location     : num  8 NA 9 9 NA 10 8 10 NA 10 ...
##  $ review_scores_value        : num  9 NA 9 9 NA 10 7 10 NA 9 ...
##  $ license                    : chr  "Exempt" "" "" "Exempt" ...
##  $ is_instant_bookable        : chr  "true" "true" "true" "true" ...
##  $ reviews_per_month          : num  0.55 NA 1.17 3.41 NA 7.72 1.41 1.11 NA 0.84 ...
##  $ country                    : chr  "spain" "spain" "spain" "spain" ...
##  $ city                       : chr  "malaga" "madrid" "mallorca" "barcelona" ...
##  $ insert_date                : chr  "2020-02-29" "2019-10-16" "2020-04-23" "2018-10-10" ...
```

Com podem observar existeix un gran nombre de variables i de diferents tipus. Trobem des de descripcions obertes dels allotjaments, comentaris dels usuaris, valoracions.. fins a les característiques de cada allotjament com ara el nombre d'habitacions, el preu, número mínim i màxim de nits entre d'altres.

Per tant, es disposa de suficient informació que tot seguit netejarem per evitar les variables que no aporten valor estadístic a l'estudi, com ara els comentaris oberts dels usuaris entre altres variables que quedaran fora de l'anàlisi per facilitar la comprensió.

L'objectiu principi de l'estudi és mostrar quines són les característiques més habituals dels allotjaments turístics a Espanya juntament amb les variables més importants que determinen el preu/nit final per tal de predir la variable objectiu preu/nit segons les característiques de l'allotjament.

# Neteja de dades

El primer pas és definir les variables del joc de dades original que utilitzarem per a l'anàlisi.

## Selecció de variables

```
# Selecció de variables del joc de dades original
data <- data[, c(11,12,13,14,15,17,18,19,25,28,29,30,31,32,33,34,39)]
```

## Tipus de variables

```
# Resum de les variables
str(data)
```

```
## 'data.frame':    10000 obs. of  17 variables:
##  $ room_type                  : chr  "Entire home/apt" "Entire home/apt" "Entire home/apt" "Private r
##  $ accommodates               : int  5 4 2 2 3 2 6 6 5 7 ...
##  $ bathrooms                  : int  1 1 1 1 1 1 2 1 1 2 ...
##  $ bedrooms                   : int  3 1 1 1 1 1 3 3 NA 3 ...
##  $ beds                       : int  4 2 2 1 2 1 5 4 3 6 ...
##  $ price                      : num  80 500 60 48 99 25 103 100 33 99 ...
##  $ minimum_nights             : int  2 1 2 1 6 2 1 4 3 2 ...
##  $ maximum_nights             : int  30 1125 15 12 31 1125 1125 28 380 60 ...
##  $ number_of_reviews          : int  11 0 15 173 0 221 3 47 0 4 ...
##  $ review_scores_rating       : num  96 NA 96 91 NA 96 73 100 NA 85 ...
##  $ review_scores_accuracy     : num  10 NA 10 9 NA 10 8 10 NA 9 ...
##  $ review_scores_cleanliness  : num  10 NA 9 10 NA 10 7 10 NA 10 ...
##  $ review_scores_checkin      : num  10 NA 10 9 NA 10 9 10 NA 10 ...
##  $ review_scores_communication: num  10 NA 10 9 NA 10 7 10 NA 9 ...
##  $ review_scores_location     : num  8 NA 9 9 NA 10 8 10 NA 10 ...
##  $ review_scores_value        : num  9 NA 9 9 NA 10 7 10 NA 9 ...
##  $ city                       : chr  "malaga" "madrid" "mallorca" "barcelona" ...
```

Definició de les variables escollides:

```
room_type: tipus caràcter que indica el tipus d'allotjament
accommodates: tipus integer amb el nombre màxim de persones per allotjament
```

bathroom: tipus integer amb el nombre de banys de l'allotjament
bedrooms: tipus integer amb el nombre d'habitacions de l'allotjament
beds: tipus integer amb el nombre de llits de l'allotjament
price: tipus numèric del preu per nit de l'allotjament
minimum_nights: tipus integer amb el nombre mínim de nits
maximum_nights: tipus integer amb el nombre màxim de nits
number_of_reviews: tipus integer amb el nombre de comentaris de l'allotjament
review_scores_rating: tipus numèric amb la puntuació dels usuaris sobre l'allotjament
review_scores_accuracy: tipus numèric amb la precisió dels detalls de l'allotjament
review_scores_cleanliness: tipus numèric amb les condicions higièniques de l'allotjament
review_scores_checkin: tipus numèric amb la facilitat del moment d'entrada al allotjament
review_scores_communication: tipus numèric que valora la comunicació del propietari amb els usuaris
review_scores_location: tipus numèric que valora la zona on està situat l'allotjament
review_scores_value: tipus numèric que puntua sobre l'avaluació de l'allotjament turístic
city: tipus caràcter que indica la ciutat on es troba el allotjament

https://datamarket.es/#alojamientos-turisticos-dataset

## Resum de les variables

```
# Resum estadístic de les variables
summary(data)
```

```
##   room_type          accommodates      bathrooms         bedrooms
##  Length:10000       Min.   : 1.000   Min.   : 0.000   Min.   : 0.000
##  Class :character   1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 1.000
##  Mode  :character   Median : 4.000   Median : 1.000   Median : 1.000
##                     Mean   : 4.256   Mean   : 1.571   Mean   : 1.918
##                     3rd Qu.: 6.000   3rd Qu.: 2.000   3rd Qu.: 3.000
##                     Max.   :27.000   Max.   :22.000   Max.   :16.000
##                                      NA's   :78       NA's   :75
##       beds            price        minimum_nights     maximum_nights
##  Min.   : 0.000   Min.   :   9.0   Min.   :   1.000   Min.   :1.000e+00
##  1st Qu.: 1.000   1st Qu.:  45.0   1st Qu.:   1.000   1st Qu.:9.000e+01
##  Median : 2.000   Median :  76.0   Median :   2.000   Median :1.125e+03
##  Mean   : 2.913   Mean   : 129.5   Mean   :   5.067   Mean   :2.155e+05
##  3rd Qu.: 4.000   3rd Qu.: 130.0   3rd Qu.:   4.000   3rd Qu.:1.125e+03
##  Max.   :40.000   Max.   :9949.0   Max.   :1000.000   Max.   :2.147e+09
##  NA's   :65
##  number_of_reviews review_scores_rating review_scores_accuracy
##  Min.   :  0.00    Min.   : 20.00       Min.   : 2.000
##  1st Qu.:  0.00    1st Qu.: 89.00       1st Qu.: 9.000
##  Median :  5.00    Median : 94.00       Median :10.000
##  Mean   : 24.84    Mean   : 91.76       Mean   : 9.438
##  3rd Qu.: 25.00    3rd Qu.: 98.00       3rd Qu.:10.000
##  Max.   :535.00    Max.   :100.00       Max.   :10.000
##                    NA's   :2782         NA's   :2793
##  review_scores_cleanliness review_scores_checkin review_scores_communication
##  Min.   : 2.000            Min.   : 2.000        Min.   : 2.000
##  1st Qu.: 9.000            1st Qu.: 9.000        1st Qu.: 9.000
##  Median :10.000            Median :10.000        Median :10.000
##  Mean   : 9.283            Mean   : 9.596        Mean   : 9.591
```

```
##  3rd Qu.:10.000           3rd Qu.:10.000        3rd Qu.:10.000
##  Max.   :10.000           Max.   :10.000        Max.   :10.000
##  NA's   :2788             NA's   :2793          NA's   :2792
##  review_scores_location review_scores_value     city
##  Min.   : 2.000         Min.   : 2.000      Length:10000
##  1st Qu.: 9.000         1st Qu.: 9.000      Class :character
##  Median :10.000         Median : 9.000      Mode  :character
##  Mean   : 9.523         Mean   : 9.097
##  3rd Qu.:10.000         3rd Qu.:10.000
##  Max.   :10.000         Max.   :10.000
##  NA's   :2792           NA's   :2792
```

De manera general observem com disposem d'informació d'allotjaments turístics amb grans diferències en les característiques. Existeixen allotjaments on només pot residir 1 usuari i altres fins a 27, encara que la mitjana es troba en 4 persones i, per tant, parlarem sobretot d'allotjaments de dimensions estàndards. Les característiques d'un allotjament estàndard són la disponibilitat d'entre 1 i 2 banys (1,5 de mitjana), 2 habitacions i 3 llits, encara que existeixen allotjaments amb 22 banys (possiblement ja es detecta algun outlier), 16 habitacions o 40 llits. El preu per nit es troba de mitjana en els 129€, encara que la mediana dels allotjaments analitzats es troba per sota dels 76€/nit.

Pel que fa a nombre de nits mínim, observem com la mitjana es troba en 5 nits, però la gran majoria es tracta d'allotjaments d'una o dues nits mínimes de reserva. Pel que fa al nombre màxim de nits es detecta un possible error en les dades que més endavant caldrà corregir.

El nombre de reviews d'usuaris destacar que mínim el 25% d'allotjaments no es disposa de cap, tot i així la mitjana es de gairebé 25 comentaris per allotjament, amb un màxim de 535.

La variable review_scores_rating valora de 0 a 100 l'opinió donada pels usuaris sobre l'allotjament, mentre que la resta de variables de puntuació són sobre un màxim de 10. Observem com la majora de mitjanes de valoració es troba al voltant del 9 encara que es disposa de moltes reviews sense valoració informada que més endavant corregirem.

Finalment, queda el camp city on s'informa de la ciutat on es troba l'allotjament.

## Eliminació de valors nuls i outliers

```
# Verifiquem nombre valors absents
missing <- data[is.na(data),]
dim(missing)
```

```
## [1] 19750    17
```

```
# Creem una copia del joc de dades original per la neteja
data_clean <- data

# Reemplacem les variables reviews per la mitjana de cada una d'elles
data_clean$review_scores_rating[is.na(data_clean$review_scores_rating)]<-round(mean(data_clean$review_s
data_clean$review_scores_accuracy[is.na(data_clean$review_scores_accuracy)]<-round(mean(data_clean$revi
data_clean$review_scores_cleanliness[is.na(data_clean$review_scores_cleanliness)]<-round(mean(data_clea
data_clean$review_scores_checkin[is.na(data_clean$ review_scores_checkin)]<-round(mean(data_clean$ revi
data_clean$review_scores_communication[is.na(data_clean$review_scores_communication)]<-round(mean(data_
data_clean$review_scores_location[is.na(data_clean$review_scores_location)]<-round(mean(data_clean$revi
data_clean$review_scores_value[is.na(data_clean$review_scores_value)]<-round(mean(data_clean$review_sco
```

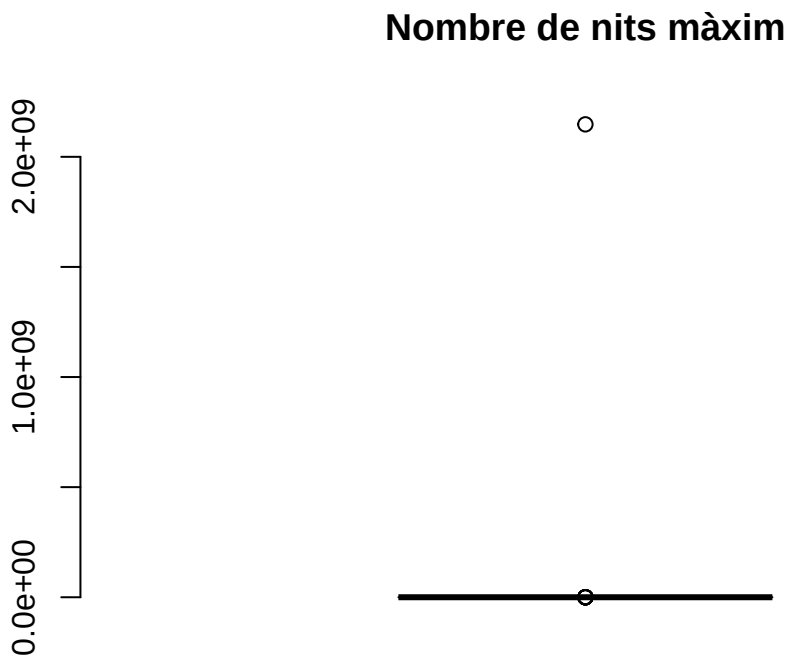
```r
# Reemplacem les variables bathrooms, bedrooms i beds amb la moda de cada variable
# Es crea la funció per extreure la moda
Mode <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]}

# S'aplica la moda als valors perduts de les variables
data_clean$bathrooms[is.na(data_clean$bathrooms)] <- Mode(data_clean$bathrooms)
data_clean$bedrooms[is.na(data_clean$bedrooms)] <- Mode(data_clean$bedrooms)
data_clean$beds[is.na(data_clean$beds)] <- Mode(data_clean$beds)
```

Per tractar els valors absents s'ha utilitzat diferents mètodes per omplir els gairebé 2.800 registres sense valors. Per donar un valor a les reviews s'usa la mitjana de cada variable independent de la resta de variables. Com s'ha estudiat anteriorment, les mitjanes de valoracions es troben entre 9 i 9,5 que al treballar sense nombres decimals, els valors buits seran substituïts amb un 9 o amb un 10 segons la mitjana de cada variable.

Pel tractament de les variables bathrooms, bedrooms i beds, utilitzarem la moda, valor que més vegades es repeteix en cada una de les variables.

```r
# Identifiquem outliers i apliquem diferents criteris

# En el cas del nombre màxim de nits s'eliminen els outliers al ser pocs registres i considerar-se com
box_MaxNights <- boxplot(data_clean$maximum_nights, col="skyblue", frame.plot=F, tittle ="MaxNights", ma
```

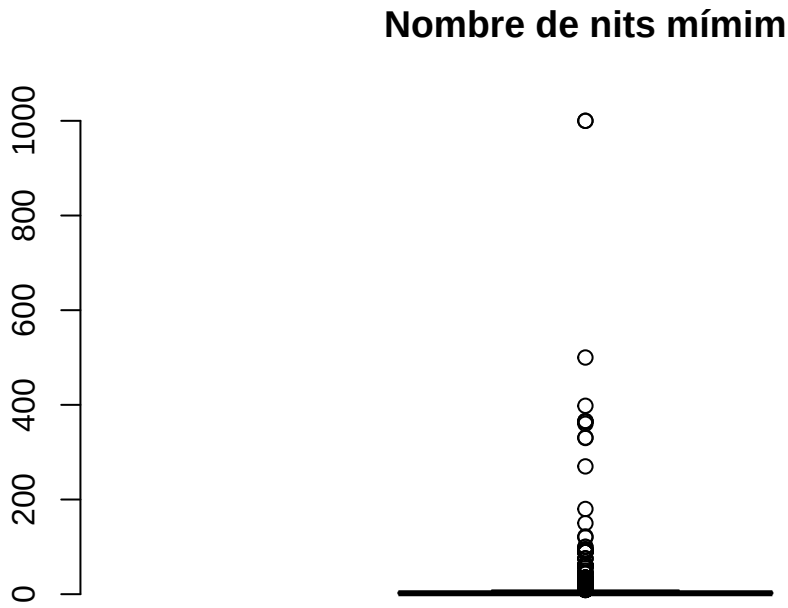
```
## Warning in x[floor(d)] + x[ceiling(d)]: NAs producidos por enteros excedidos
```



```
box_MaxNights$out
```

```
## [1]       30000      142365       29999 2147483647       10000       10000        5000
## [8]        5000
```

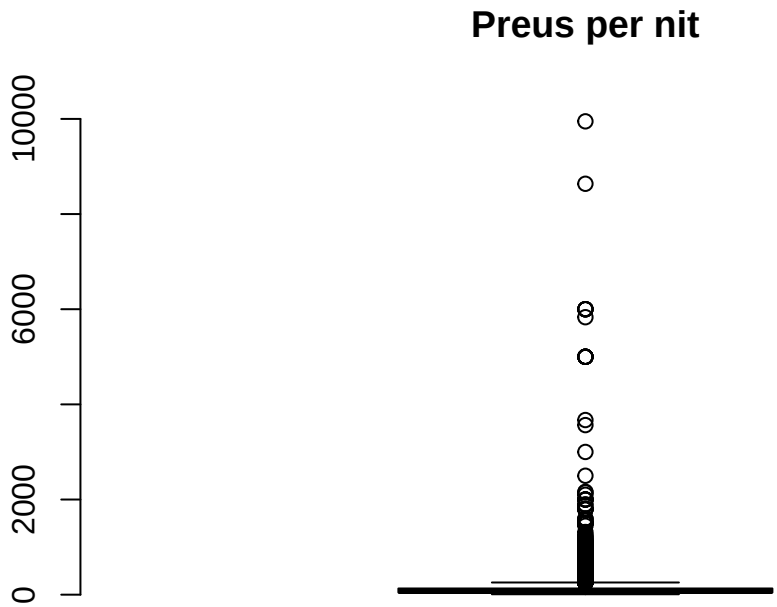
```
data_clean<-data_clean[!(data_clean$maximum_nights %in% box_MaxNights$out),]

# En el cas del nombre mínim de nits s'eliminen els outliers superiors a 60 nits
box_MinNights <- boxplot(data_clean$minimum_nights, col="skyblue", frame.plot=F, main="Nombre de nits mi
```




```
sum(data_clean$minimum_nights>60)
```

```
## [1] 44
```

```
data_clean <- data_clean[!(data_clean$minimum_nights > 60),]

# En el cas del preu eliminem els registres superiors a 4000
box_Price <- boxplot(data_clean$price, col="skyblue", frame.plot=F, main="Preus per nit")
```




```
sum(data_clean$price>4000)
```

```
## [1] 16
```

```
data_clean <- data_clean[!(data_clean$price > 4000),]

# Revisem el nombre de registres que seran analitzats
nrow(data_clean)
```

```
## [1] 9932
```

Un cop revisats els valors màxims de les variables numèriques del joc de dades es revisa que les variables maximum_nights, minimum_nights i price són les que tenen més probabilitats de contenir valors outliers. En els diagrames de caixes observem com és així, i es determinen diferents procediments per eliminar-los. Encara que no és del tot recomanable eliminar registres per possibles outliers, de l'estudi es detecta que es tracta d'errors com és en el cas del nombre màxim de nits o de valors fora de les característiques estàndards dels allotjaments a analitzar.

Per tant, es decideix eliminar els outliers de les variables maximum_nights, a més d'aquells registres el qual tenen un nombre mínim de 60 nits o que tenen un preu per nit major a 4.000€.

El joc de dades finalment contindrà 9.932 registres dels inicialment 10.000.

```r
# Es revisa que tots els camps es troben nets, sense valors perduts i cap outlier destacat.
summary(data_clean)
```

```
##   room_type          accommodates      bathrooms         bedrooms
##  Length:9932        Min.   : 1.000   Min.   : 0.000   Min.   : 0.000
##  Class :character   1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 1.000
##  Mode  :character   Median : 4.000   Median : 1.000   Median : 1.000
##                     Mean   : 4.264   Mean   : 1.568   Mean   : 1.913
##                     3rd Qu.: 6.000   3rd Qu.: 2.000   3rd Qu.: 3.000
##                     Max.   :27.000   Max.   :22.000   Max.   :16.000
##       beds            price        minimum_nights   maximum_nights
##  Min.   : 0.000   Min.   :   9.0   Min.   : 1.000   Min.   :   1.0
##  1st Qu.: 1.000   1st Qu.:  46.0   1st Qu.: 1.000   1st Qu.:  71.5
##  Median : 2.000   Median :  76.0   Median : 2.000   Median :1125.0
##  Mean   : 2.906   Mean   : 120.4   Mean   : 4.229   Mean   : 762.7
##  3rd Qu.: 4.000   3rd Qu.: 130.0   3rd Qu.: 4.000   3rd Qu.:1125.0
##  Max.   :40.000   Max.   :3671.0   Max.   :60.000   Max.   :1125.0
##  number_of_reviews review_scores_rating review_scores_accuracy
##  Min.   :  0.00    Min.   : 20.00       Min.   : 2.000
##  1st Qu.:  0.00    1st Qu.: 91.00       1st Qu.: 9.000
##  Median :  5.00    Median : 92.00       Median : 9.000
##  Mean   : 24.93    Mean   : 91.84       Mean   : 9.317
##  3rd Qu.: 26.00    3rd Qu.: 97.00       3rd Qu.:10.000
##  Max.   :535.00    Max.   :100.00       Max.   :10.000
##  review_scores_cleanliness review_scores_checkin review_scores_communication
##  Min.   : 2.000            Min.   : 2.000        Min.   : 2.000
##  1st Qu.: 9.000            1st Qu.:10.000        1st Qu.:10.000
##  Median : 9.000            Median :10.000        Median :10.000
##  Mean   : 9.205            Mean   : 9.708        Mean   : 9.706
##  3rd Qu.:10.000            3rd Qu.:10.000        3rd Qu.:10.000
##  Max.   :10.000            Max.   :10.000        Max.   :10.000
##  review_scores_location review_scores_value     city
##  Min.   : 2.000         Min.   : 2.000      Length:9932
##  1st Qu.: 9.000         1st Qu.: 9.000      Class :character
##  Median :10.000         Median : 9.000      Mode  :character
##  Mean   : 9.656         Mean   : 9.071
##  3rd Qu.:10.000         3rd Qu.:10.000
##  Max.   :10.000         Max.   :10.000
```

```r
# Exportació del joc de dades net
write.csv(data_clean,"./clean_alojamientos-turisticos-sample.csv", row.names = FALSE)
```

## Estudi de les distribucions i correlacions

A continuació mostrem gràficament les dades més importants del conjunt de dades un cop net.

```r
# Carreguem les llibreries necessàries
if (!require('ggplot2')) install.packages('ggplot2'); library('ggplot2')
```

```
## Loading required package: ggplot2
```

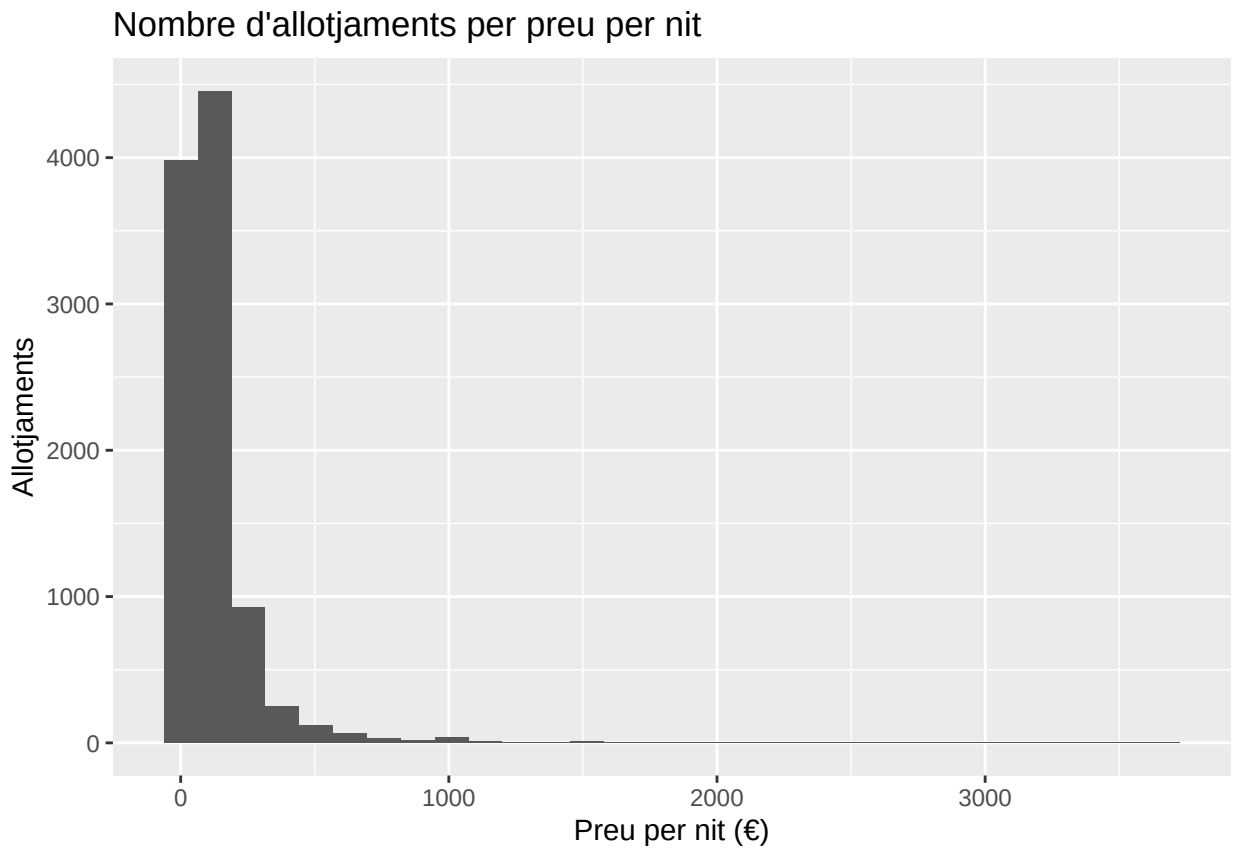
```r
if (!require('dplyr')) install.packages('dplyr'); library('dplyr')
```

```
## Loading required package: dplyr

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
if (!require('ggcorrplot')) install.packages('ggcorrplot'); library('ggcorrplot')
```

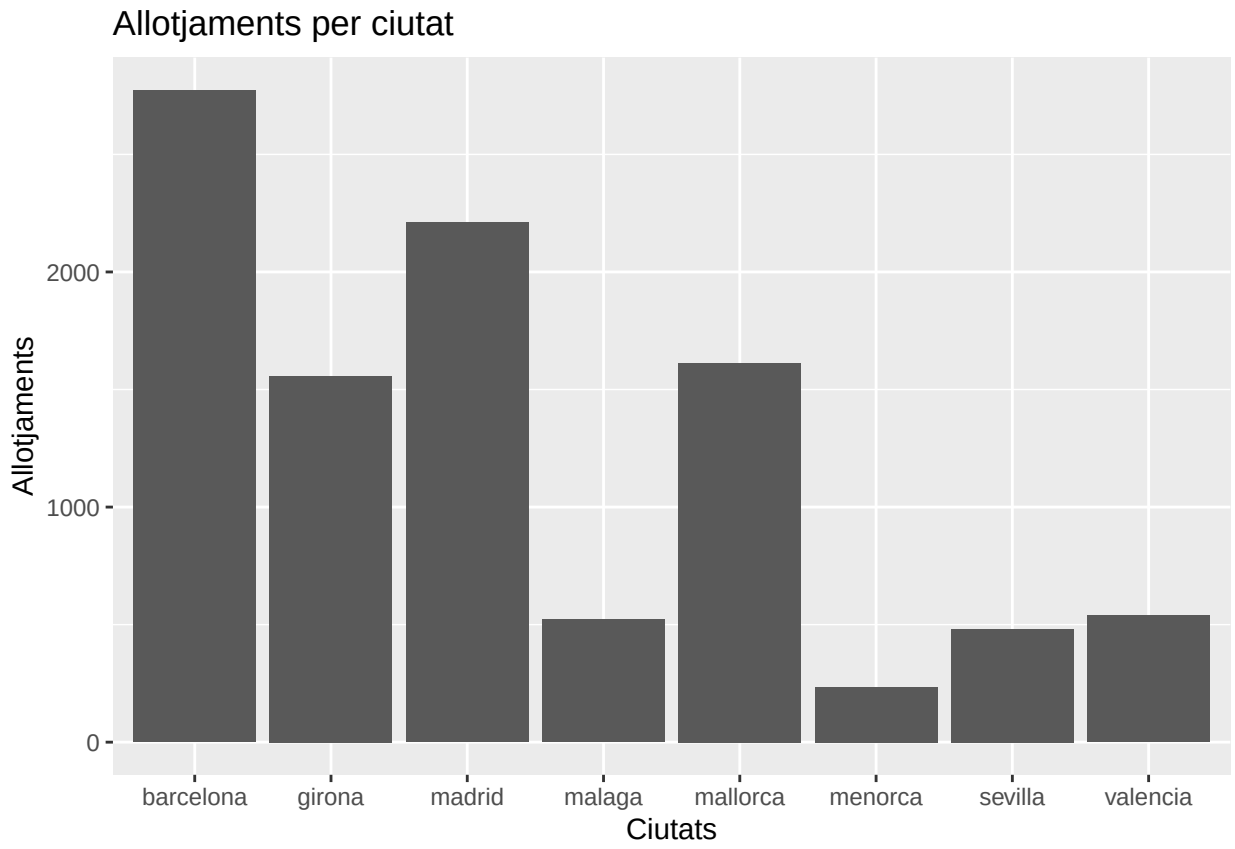
```
## Loading required package: ggcorrplot
```

```r
# Histograma de preus
histobyPrice<-ggplot(data_clean,aes(price)) + geom_histogram(bins=30) + labs(x="Preu per nit (€)", y="Al
  scale_fill_manual(values=c("blue","#008000")) + ggtitle("Nombre d'allotjaments per preu per nit")
histobyPrice
```



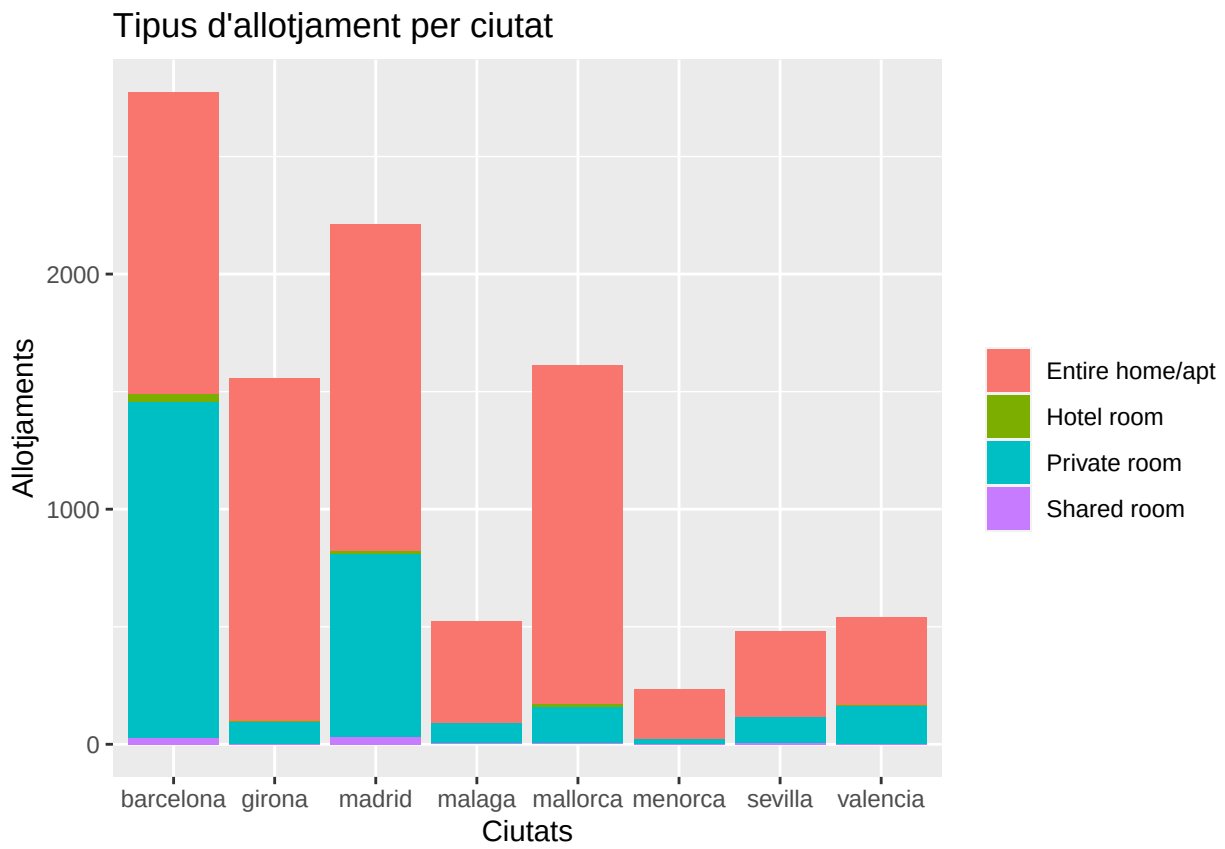

Observem com la gran majoria d'allotjaments es situen per sota dels 200€/nit.

```
# Gràfic per ciutats
plotbyCity <- ggplot(data_clean,aes(city)) +geom_bar() + labs(x="Ciutats", y="Allotjaments") + guides(fi
  scale_fill_manual(values=c("blue","#008000")) +ggtitle("Allotjaments per ciutat")
plotbyCity
```




Les ciutats amb més ofertes d'allotjament són Barcelona, seguit de Madrid, Mallorca i Girona.

```
#plotbyRisc<-ggplot(data_Risc,aes(default))+geom_bar() + labs(x="Risc", y="Clients") + guides(fill=guid
  #scale_fill_manual(values=c("blue","#008000"))  +ggtitle("Risc")

plotbyRoomCity <- ggplot(data_clean,aes(city,fill=room_type)) +geom_bar() + labs(x="Ciutats", y="Allotj
  ggtitle("Tipus d'allotjament per ciutat")
plotbyRoomCity
```




La gran majoria d'allotjaments disponibles són d'apartaments sencers, encara que també existeix força oferta d'habitacions privades. En canvi d'habitcaions d'hotel o d'habitacions compartides no existeix un gran nombre d'ofertes en la plataforma Airbnb.

```
# Mostrem gràficament les correlacions existents entre les característiques numèriques del joc de dades
data_num <- dplyr::select_if(data_clean, is.numeric)
r <- cor(data_num, use="complete.obs")
ggcorrplot(r,hc.order = TRUE, type = "lower")
```


Finalment, elaborem un primer anàlisi gràfic de correlacions per poder començar a predir quines variables afecten en major mesura al preu.

Com hem comentat anteriorment, totes les reviews mostraven uns valors força elevats i semblants, per tant no determinen el valor del preu. En canvi, si que observem una correlació positiva amb el nombre d'habitacions, llits i nombre de persones.

# Normalització

## Revisió de dades normalitzades

Un cop tenim les dades necessàries per a l'estudi, comprovarem si la variable a analitzar 'price' segueix una distribució normal. Per això, realitzarem una inspecció visual mitjançant els gràfics quantile-quantile plot i l'histograma, i un contrast de normalitat de Lilliefors.

```
#Carreguem les llibreries
if (!require('kableExtra')) install.packages('kableExtra'); library('kableExtra')
```

```
## Loading required package: kableExtra
```

```
## Warning: package 'kableExtra' was built under R version 4.1.2
```

```
##
## Attaching package: 'kableExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
if (!require('knitr')) install.packages('knitr'); library('knitr')
```

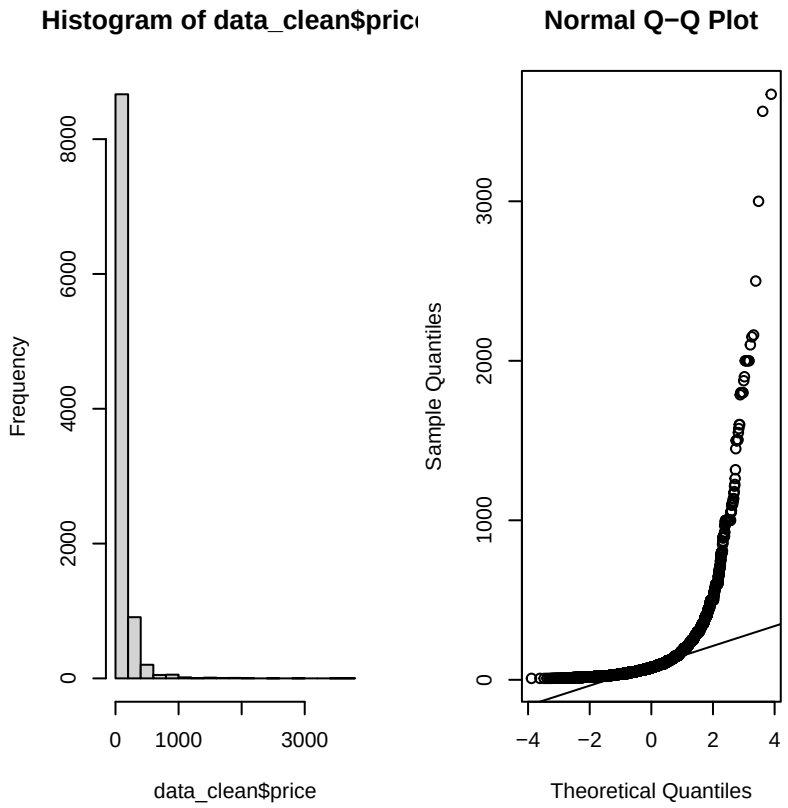
```
## Loading required package: knitr
```

```
if (!require('nortest')) install.packages('nortest'); library('nortest')
```

```
## Loading required package: nortest
```

```
#Gràfics
par(mfrow=c(1,3))
#Histograma
hist(data_clean$price)
qqnorm(data_clean$price)
#Gràfic quantile
qqline(data_clean$price)
#Costrast de normalitat
lillie.test(data_clean$price)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data_clean$price
## D = 0.25696, p-value < 2.2e-16
```




Com podem veure en el gràfic Q-Q, els punts de la mostra no es representen sobre la distribució teòrica, de manera que no podríem assumir que té una distribució normal. Per altre banda, en el test de Lilliefors s'obté un pvalue molt petit, de manera que no podem afirmar que la variable 'price' segueixi una distribució normal.

Per tant, no podem afirmar que la 'price' es distribueix segons una distribució normal.

## Transformació de dades normalitzades

A pesar que la variable 'price' no segueix una distribució normal, no realitzarem cap normalització en la variable, ja que volem treballar amb els valors reals.

## Aplicació de proves estadístiques

### Contrast d'hipòtesis de dues mostres

Es pressuposa que les habitacions dels hotels acostumen a ser més cares que les habitacions privades i els apartaments. De manera que, volem conèixer si hi ha diferències entre els preus de les habitacions dels hotels i els apartaments, i les habitacions privades. En aquest cas, aplicarem un test d'hipòtesi de dues mostres sobre la mitjana, ja que tenim una mostra de grandària gran(>30), de manera que pel teorema del límit central, assumim normalitat. Com que no coneixem la variància de la població, aplicarem una distribució de t.

```
#Seleccionem els preu del conjunt de dades del tipus habitació d'hotel
hotel_room <- data_clean$price[data_clean$room_type=="Hotel room"]
```

```r
#Seleccionem els preu del conjunt de dades del tipus habitació privada
private_room <- data_clean$price[data_clean$room_type=="Private room"]
#Seleccionem els preu del conjunt de dades del tipus habitació apartament
apartment <- data_clean$price[data_clean$room_type=="Entire home/apt"]

#Mitjana del preu de les habitacions dels hotels
mean(hotel_room)
```

```
## [1] 266.1955
```

```r
#Mitjana del preu de les habitacions privades
mean(private_room)
```

```
## [1] 60.90966
```

```r
#Mitjana del preu dels apartaments
mean(apartment)
```

```
## [1] 143.81
```

Observant la mitja de preus, fa pensar que el preu de les habitacions dels hotels són més cares que les habitacions privades i els apartaments.

```r
#Apliquem el test d'hipòtesi de dues mostres sobre la mitjana de preu les habitacions dels hotels i les
t.test(hotel_room, private_room, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  hotel_room and private_room
## t = 4.7294, df = 64.369, p-value = 6.358e-06
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  132.8464      Inf
## sample estimates:
## mean of x mean of y
## 266.19554  60.90966
```

```r
#Apliquem el test d'hipòtesi de dues mostres sobre la mitjana de preu les habitacions dels hotels i els
t.test(hotel_room, apartment, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  hotel_room and apartment
## t = 2.8204, df = 64.293, p-value = 0.003186
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  49.96603      Inf
## sample estimates:
## mean of x mean of y
##  266.1955  143.8100
```

Observem que els valors de p són menors que el valor de significació fixat(0.05), de manera que rebutgem la hipòtesi nul · la. Per tant, podem afirmar que els preus de les habitacions dels hotels són més cares que les habitacions privades i els apartaments.

### Correlació entre variables

Ara realitzarem un anàlisi de correlació entre les diferents variables per determinar quines variables estan més correlacionades amb el preu de les habitacions. Per això, durem a terme una correlació de Person.

```
#Coeficients de corrlació de Pearson
cor(data_clean[,c("price", "accommodates", "bathrooms", "bedrooms", "beds", "review_scores_rating", "re
```

```
##                              price accommodates    bathrooms     bedrooms
## price                   1.00000000   0.42811872  0.410505071  0.422972892
## accommodates            0.42811872   1.00000000  0.637933313  0.852718218
## bathrooms               0.41050507   0.63793331  1.000000000  0.680049962
## bedrooms                0.42297289   0.85271822  0.680049962  1.000000000
## beds                    0.36367205   0.84999873  0.624477334  0.799766569
## review_scores_rating    0.01010125  -0.01655891  0.033286003  0.014511125
## review_scores_location  0.06140475  -0.01548084  0.034371106 -0.006308818
## review_scores_value    -0.02713514  -0.05777971 -0.003002904 -0.029629185
##                                beds review_scores_rating review_scores_location
## price                   0.36367205           0.01010125            0.061404748
## accommodates            0.84999873          -0.01655891           -0.015480838
## bathrooms               0.62447733           0.03328600            0.034371106
## bedrooms                0.79976657           0.01451112           -0.006308818
## beds                    1.00000000          -0.01079989           -0.012957782
## review_scores_rating   -0.01079989           1.00000000            0.472734630
## review_scores_location -0.01295778           0.47273463            1.000000000
## review_scores_value    -0.05195337           0.78260717            0.466116810
##                        review_scores_value
## price                       -0.027135142
## accommodates                -0.057779714
## bathrooms                   -0.003002904
## bedrooms                    -0.029629185
## beds                        -0.051953369
## review_scores_rating         0.782607172
## review_scores_location       0.466116810
## review_scores_value          1.000000000
```

A partir dels resultats obtinguts, podem veure que les variables: 'accomodates', 'bedrooms', 'bathrooms' són les variables que estan més correlacionades amb la variable 'price'. Té sentit que les habitacions més cares estiguin relacionades amb el nombre de persones que es poden allotjar, el nombre d'habitacions i el nombre de banys que té l'allotjament.

Tanmateix, no podem considerar que hi hagi una correlació forta, ja que els coeficients de correlació són de 0.4 una correlació moderada-baixa.

### Model de regressió lineal

Un dels objectius de l'estudi és conèixer quins són característiques influeixen en els preus dels allotjaments de Airbnb. Per això, calcularem un model de regressió lineal per tal de conèixer l'equació del model que ens permeti conèixer el preu de l'allotjament a partir de les seves característiques.

Per tal d'obtenir un model eficient, utilitzarem diferents models amb les variables més correlacionades amb el preu a partir dels resultats obtinguts en en la correlació de variables. Per tal d'avaluar la bondat de l'ajust del model utilizarem el coeficient de determinació (R2).

```
#Creació del model a partir de la variable 'accomodates'
model_1<- lm(price~accommodates, data=data_clean)

#Resum del model
summary(model_1)
```

```
##
## Call:
## lm(formula = price ~ accommodates, data = data_clean)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -427.4  -51.9  -20.3    7.1 3448.1
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.4437     2.9050   1.185    0.236
## accommodates  27.4341     0.5811  47.207   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 151.1 on 9930 degrees of freedom
## Multiple R-squared:  0.1833, Adjusted R-squared:  0.1832
## F-statistic:  2228 on 1 and 9930 DF,  p-value: < 2.2e-16
```

Podem veure que amb la variable 'accomodates' és significativa amb un p-valor de 2.2e-16 existint una relació lineal positiva entre totes dues variables, amb un coeficient de determinació ajustat de 0.1833. Malgrat això, el coeficient de determinació és força baix per explicar el model.

```
#Creació del model afegint la variable 'bedrooms'
model_2<- lm(price~accommodates+bedrooms, data=data_clean)

#Resum del model
summary(model_2)
```

```
##
## Call:
## lm(formula = price ~ accommodates + bedrooms, data = data_clean)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -476.3  -48.6  -21.1    7.4 3434.0
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.2394     2.8949   0.083    0.934
## accommodates  15.8380     1.1042  14.344   <2e-16 ***
## bedrooms      27.5192     2.2344  12.316   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 149.9 on 9929 degrees of freedom
## Multiple R-squared:  0.1956, Adjusted R-squared:  0.1954
## F-statistic:  1207 on 2 and 9929 DF,  p-value: < 2.2e-16
```

Afegint la variable 'bedrooms' al model, s'observa que la variable és significativa i s'aconsegueix millorar el coeficient de determinació (0.1956) però segueix insuficient per explicar correctament el model.

```
#Creació del model afegint la variable 'bathrooms'
model_3<- lm(price~accommodates+bedrooms+bathrooms, data=data_clean)

#Resum del model
summary(model_3)
```

```
##
## Call:
## lm(formula = price ~ accommodates + bedrooms + bathrooms, data = data_clean)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -748.2  -48.3  -20.4    8.7 3386.9
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -17.266      3.037  -5.685 1.35e-08 ***
## accommodates   13.022      1.101  11.822  < 2e-16 ***
## bedrooms       14.158      2.341   6.047 1.53e-09 ***
## bathrooms      35.132      2.082  16.872  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 147.8 on 9928 degrees of freedom
## Multiple R-squared:  0.218,  Adjusted R-squared:  0.2178
## F-statistic: 922.5 on 3 and 9928 DF,  p-value: < 2.2e-16
```

Afegint la tercera variable amb més correlació amb el preu 'bedrooms' al model, veiem que la variable és significativa i el coeficient de determinació millora (0.218) però no el suficient per considerar el model com a acceptable.

```
#Creació del model a partir de la variable 'city'
model_4<- lm(price~factor(city), data=data_clean)

#Resum del model
summary(model_4)
```

```
##
## Call:
## lm(formula = price ~ factor(city), data = data_clean)
##
## Residuals:
##    Min     1Q Median     3Q    Max
```

```
## -189.1  -64.1  -36.3    8.8 3472.9
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            97.228      3.086  31.509  < 2e-16 ***
## factor(city)girona     38.064      5.145   7.398 1.49e-13 ***
## factor(city)madrid      1.786      4.633   0.386  0.69985
## factor(city)malaga    -13.082      7.746  -1.689  0.09128 .
## factor(city)mallorca  100.886      5.088  19.827  < 2e-16 ***
## factor(city)menorca    85.662     11.040   7.759 9.38e-15 ***
## factor(city)sevilla     6.257      8.026   0.780  0.43567
## factor(city)valencia  -21.988      7.649  -2.875  0.00405 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 162.5 on 9924 degrees of freedom
## Multiple R-squared:  0.05561,    Adjusted R-squared:  0.05494
## F-statistic: 83.48 on 7 and 9924 DF,  p-value: < 2.2e-16
```

En aquest cas, volem explicar el preu a partir de la variable 'city'. Podem veure com el model és significatiu amb un p_valor de 2.2e-16 existint una relació lineal negativa entre les ciutats de Girona, Màlaga i València. Això no obstant, el coeficient de determinació és molt petit (0.05), de manera que podem considerar el model força pobre.

```
#Creació del model afegint la variable 'city'
model_5<- lm(price~accommodates+bedrooms+bathrooms+factor(city), data=data_clean)

#Resum del model
summary(model_5)
```

```
##
## Call:
## lm(formula = price ~ accommodates + bedrooms + bathrooms + factor(city),
##     data = data_clean)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -685.2  -47.2  -21.1    9.7 3376.3
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -13.316      3.682  -3.616  0.00030 ***
## accommodates           14.086      1.127  12.501  < 2e-16 ***
## bedrooms               12.802      2.384   5.369 8.09e-08 ***
## bathrooms              32.746      2.100  15.590  < 2e-16 ***
## factor(city)girona    -16.853      4.889  -3.447  0.00057 ***
## factor(city)madrid      5.721      4.222   1.355  0.17544
## factor(city)malaga    -28.631      7.080  -4.044 5.29e-05 ***
## factor(city)mallorca   13.114      4.990   2.628  0.00860 **
## factor(city)menorca    21.295     10.113   2.106  0.03525 *
## factor(city)sevilla    -3.094      7.314  -0.423  0.67229
## factor(city)valencia  -32.402      6.940  -4.669 3.07e-06 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 147.3 on 9921 degrees of freedom
## Multiple R-squared:  0.2245, Adjusted R-squared:  0.2237
## F-statistic: 287.2 on 10 and 9921 DF,  p-value: < 2.2e-16
```

Podem veure que en afegir la variable 'city' al model, el coeficient de determinació millora (0.2245) però era d'esperar tenint en compte la bondat de l'ajust entre la variable 'price' i 'city' en el model anterior.

Per tant, no podem considerar el model prou consistent per a explicar el preu a partir d'aquestes variables.

## Predicció del model

Per tal de comprovar la validesa del model, realitzarem la predicció del model comprovant els valors predits amb els valors reals. Crearem un conjunt d'entrenament per crear el model i un conjunt de prova per predir el preu de l'allotjament a partir del model que hem creat amb el conjunt d'entrenament.

```
#Mida del conjunt d'entreenament
ntrain <- nrow(data_clean)*0.8

#Mida del conjunt de prova
ntest <- nrow(data_clean)*0.2
set.seed(1)

#Seleccionem elements de manera aleatoria del conjunt de dades
index_train<-sample(1:nrow(data_clean),size = ntrain)

#Conjunt d'entreenament
train<-data_clean[index_train,]

#Conjunt de proba
test<-data_clean[-index_train,]

#Creació del model a partir del conjunt d'entrenament
model<-lm(price ~ accommodates + bedrooms + bathrooms + factor(city), data=train)

#Resum del model
summary(model)
```

```
##
## Call:
## lm(formula = price ~ accommodates + bedrooms + bathrooms + factor(city),
##     data = train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -610.8  -48.1  -22.3    9.8 3381.5
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -10.578      4.193  -2.523 0.011659 *
## accommodates           14.021      1.307  10.725  < 2e-16 ***
## bedrooms               14.827      2.802   5.292 1.24e-07 ***
```

```
## bathrooms              29.070      2.357  12.332  < 2e-16 ***
## factor(city)girona    -18.749      5.582  -3.359 0.000787 ***
## factor(city)madrid      7.851      4.838   1.623 0.104663
## factor(city)malaga    -28.432      8.256  -3.444 0.000577 ***
## factor(city)mallorca   12.312      5.692   2.163 0.030570 *
## factor(city)menorca     7.222     11.215   0.644 0.519629
## factor(city)sevilla    -1.963      8.230  -0.238 0.811535
## factor(city)valencia  -35.477      7.988  -4.441 9.07e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 150.6 on 7934 degrees of freedom
## Multiple R-squared:  0.2115, Adjusted R-squared:  0.2105
## F-statistic: 212.8 on 10 and 7934 DF,  p-value: < 2.2e-16
```

```
#Predicció del model
prob <- predict(model, test, type="response")

#Calcul diferència % entre valor real i valor predit
mc_sl<-data.frame(real=test$price, predicted= prob, dif=ifelse(test$price>prob, -prob*100/test$price,pr

#Taula amb els valors reals i valors predits
colnames(mc_sl)<-c("Real","Predecido","Dif%")
head(mc_sl)
```

```
##    Real Predecido      Dif%
## 1    80 104.64815 130.81018
## 11   40 169.19515 422.98787
## 13  130 145.61359 112.01046
## 17   55 118.47399 215.40725
## 20  350  69.21291 -19.77512
## 23  280 369.28587 131.88781
```

Podem veure que el model del conjunt d'entrenament, té un coeficient de determinació (R2) de 0.2115, el que provocarà que el model no ajusti del tot correctament els preus predits. A pesar d'això podem veure que les variables que hem utilitzat el model són significatives.

Observant els resultats predits, podem veure que no acaba d'ajustar correctament els preus, cosa que ja havíem comentat prèviament.

# Conclusions

Encara que aparentment extreure conclusions hagi sigut més complicat de l'esperat, podem determinar una sèrie de resultats extrets de l'anàlisi elaborat.

El fet que les variables de reviews del usuari es situessin la gran majoria per sobre del 9 ha fet que s'hagi de descartar aquest tipus de variables. El joc de dades s'ha mostrat massa homogeni el qual ha provocat no poder conèixer com afecten el preu final de l'allotjament. La correlació existent amb la variable preu no existia i, per tant, no han servit per determinar el valor de la variable objectiu.

Tot i així la resta de variables han permès extreure les següents conclusions:

- La gran majoria d'oferta d'allotjaments es troben en les ciutats més turístiques de l'Estat.

- Predominen els lloguers d'apartaments complets i el lloguer d'habitacions privades, sent les habitacions d'hotel d'un preu superior.
- Com era d'esperar inicialment, el fet que l'allotjament permeti allotjar-se més usuaris el preu s'incrementa, per tant d'igual manera, si disposa de més llits el mateix ocorrerà.
- La poca significança de les variables existents provoca que s'obtingui un model predictiu de baix coeficient de determinació.

Així que el joc de dades i les seves variables no han aconseguit respondre amb exactitud al problema inicialment plantejat. Es necessitaria estudiar si incorporar més variables, com per exemple el nivell de connexió amb transport públic o la proximitat als monuments més importants de la ciutat, millorarien el model predictiu.

## Taula de contribucions

CONTRIBUCIONS FIRMA

Investigació prèvia FBL, ORG

Redacció de les respostes FBL, ORG

Desenvolupament codi FBL, ORG