Tipologia i cicle de vida de les dades

Pràctica 2

Nom: Pedro Galán

1. Descripció del Dataset

El dataset escollit per dur a terme la pràctica correspon a dades més actuals referents a les propietats ofertes a la plataforma AirBNB ubicades a la ciutat de Barcelona. En realitat aquest dataset no prové de Kaggle, sinó del propi portal de dades obertes de AirBNB "InsideAirBNB" (http://insideairbnb.com/get-the-data.html) , on es fan públiques sota una llicencia *Creative Commons CC0 1.0 Universal (CC0 1.0) "Public Domain Dedication" license*.

La idea d'utilitzar aquest dataset si que prové de Kaggle, on hi ha un dataset similar de la ciutat de Boston, però em va semblar més interesant realitzar l'estudi sobre Barcelona per ser un entorn que ens és més proper, i perquè, a més, el dataset de Kaggle ja estava una mica tractat ja que s'havia fet una selecció prèvia de camps i així tinc la oportunitat de fer aquest pas també jo mateix i decidir els camps més interessants per l'estudi que vull fer.

Aquest dataset és important perquè conté dades tant del tipus d'habitatge ofert (característiques, ubicació, tipologia, etc.) com del propietari de la mateixa, preu al que s'ofereix i qualificació dels usuaris, així com de "popularitat" de la propietat.

La pregunta que pretenc respondre amb aquestes dades en realitat són varies i **anirien encaminades a orientar a un possible usuari de la plataforma que volgués posar una propietat en lloguer per tal de saber quin preu podria demanar i quines característiques són les que principalment hauria de tenir presents per tal d'obtenir un bon nivell d'ocupació** (tant a nivell d'equipament de l'habitatge com de serveis prestats pel hoste).

2. Neteja de les dades

2.1 Selecció de les dades d'interès a analitzar

El primer que he fet ha sigut examinar les dades i fer un breu anàlisi descriptiu de les mateixes per tal de veure els camps d'informació presents. El fitxer té molts camps que són de tipus "memo" on es descriu l'habitatge, el barri, etc. Tots aquests camps no m'interessen perquè no em permeten fer agrupacions, per tal he decidit obviar-los.

Igualment, hi ha altres camps que no aporten informació relativa a la pregunta que pretenc respondre, i de la mateixa manera també he decidit eliminar-los.

Finalment, també he eliminat identificadors redundants o que es fan servir per funcions de la plataforma, com poden ser urls auxiliars, etc.

He fet també un anàlisi del camp "Avalaibility_365" que pretenc utilitzar com indicatiu de la ocupació d'un pis. El fitxer original té un camp que es diu 'availability_365' que conté un valor entre 0 i 365, però que no ens deixa clar si són els dies que un habitatge està ocupat o està lliure. Per comprovar-ho, farem servir un segon fitxer, 'calendar.csv' obtingut de la mateixa font que conté les reserves per cada habitatge, i extreure'm la informació de dies ocupat d'aquí. Després ho compararem amb el valor del fitxer i sabrem que indica el camp.

Llegim aquest fitxer 'calendar.csv':

```
> data<-read.csv("C:/Users/Fenix/Dropbox/uoc/Master Data Science/Tipologia i cicle de vida de les dades/PRACTICA2/calendar.csv")
```

Ordenem el fitxer per id i data

```
> sorted_data<-data[order(data$listing_id, data$date),]
```

Extreiem els Id's unics de cada habitatge per obtenir un unic id per cada propietat

```
> Unique_IDs<-with(sorted_data,unique(listing_id))
>
> head(Unique_IDs)

[1] 10938 11194 18477 18653 18666 18674
```

Generem el fitxer d'ocupació per cada habitatge amb la següent comanda:

```
> Occupied<-with(sorted_data, tapply(available, listing_id, function(x) sum((x=='t'), na.rm=TRUE)))

> head(Occupied)
```

```
10938 11194 18477 18653 18666 18674
  234   349   104   131   130   181
```

Llegim ara el fitxer mestre, “listintgs.csv”:

```
> listings<-read.csv("C:/Users/Fenix/Dropbox/uoc/Master Data Science/Tipologia i cicl
e de vida de les dades/PRACTICA2/listings.csv")
```

L’ordenem també:

```
> listings<-listings[order(listings$id),]
```

I afegim la columna “occupied” com a nou camp.

```
> listings$occupied<-Occupied
```

Comparant la columna nova amb la que ja hi havia**, veiem que són iguals i que, per tant, queda clar que aquest camp mostra els total de dies ocupat per un total màxim de 365**.

Camps seleccionats

Així doncs, els camps he decidit seleccionar 32 camps com a rellevants. A continuació indico quins són, juntament amb una breu descripció de cadascun i del motiu que m’ha fet escollir-lo:

'id': Identificador únic de l’habitatge. Interessa mantenir-ho per si s’han de creuar dades posteriorment.

'host_id': Identificador únic del hoste. Interessa mantenir-ho per si s’han de creuar dades posteriorment.

'host_name': Nom del hoste. Pot ser interesant a l’hora de mostrar les dades.

'host_since': Antiguitat del hoste. Podria ser un paràmetre significatiu? Un hoste més antic té una relació directa amb propietats més reservades? (o el que seria el mateix: l’experiència és un grau?)

'host_response_time': Temps que triga un hoste a respondre. Pot ser un indicador de nivell de servei del hoste.

'host_response_rate': Rati de respostes del hoste. Un altre indicador de nivell de servei del hoste.

'host_acceptance_rate': Rati d'acceptació de reserves. Podria servir, a priori, per donar més credibilitat als resultats permetent matisar valors alts de reserves que vinguin provocats per un llindar massa baix a l'hora de denegar reserves potencialment conflictives, per exemple. O simplement per indicar si és més interesant ser restrictiu o no en aquest aspecte.

'host_listings_count': Nombre de propietats del hoste. Influeix tenir més o menys propietats en oferta perquè el rendiment de cadascuna sigui millor?

'host_identity_verified': Si la identitat del hoste està verificada. Els usuaris es fixen en aquest aspecte a l'hora de triar?

'host_has_profile_pic': Indica si el hoste té foto al perfil. És un aspecte que dona confiança a l'hora de reservar?

'neighbourhood': El barri on es troba la propietat.

'neighbourhood_cleansed': El barri però amb el camp "netejat". En teoria amb aquest camp podríem eliminar l'anterior, però abans de fer-ho prefereixo mantenir-ho per comprovar que no perdo informació.

'neighbourhood_group_cleansed': El districte al que pertany el barri.

'city': Ciutat.

'zipcode': Codi postal. Pot permetre un anàlisi geogràfic més detallat.

'property_type': Tipus de propietat.

'room_type': Tipus d'habitació.

'accommodates': capacitat.

'bathrooms': Numero de banys.

'bedrooms': Numero de habitacions.

'beds': Numero de llits.

'bed_type': Tipus de llit. Es significatiu?

'square_feet': Metres quadrats de l'habitatge.

'price': Preu al que s'ofereix.

'cleaning_fee': Taxa de neteja. Si l'habitatge té una taxa a pagar addicional per la neteja al finalitzar l'estància.

'guests_included': convidats inclosos.

'availability_365': Disponibilitat anual. Indicador de quants dies l'any està reservada la propietat.

'number_of_reviews': Numero de comentaris que té. Pot ser un indicador de popularitat.

'review_scores_rating': Rati de puntuació dels comentaris.

'cancellation_policy': Política de cancel·lació. Podria ser un apartat que tingués influència en les reserves.

'instant_bookable': Si es pot reservar al moment. Potencial indicador de nivell de servei a l'usuari.

'reviews_per_month': Comentaris mensuals. Un altre indicador de popularitat.

```
> listings<-listings[,c('id','host_id','host_name','host_since','host_response_time',
'host_response_rate',
+   'host_acceptance_rate','host_listings_count','host_identity_verified','host_has_p
rofile_pic','neighbourhood',
+   'neighbourhood_cleansed','neighbourhood_group_cleansed','city','zipcode','propert
y_type','room_type',
+   'accommodates','bathrooms','bedrooms','beds','bed_type','square_feet','price','cl
eaning_fee',
+   'guests_included','availability_365','number_of_reviews','review_scores_rating','
cancellation_policy',
+   'instant_bookable','reviews_per_month')]
```

2.2 Les dades contenen zeros o elements buits? I valors extrems? Com gestionaries cada cas?

Quant a elements buits, després de la primera anàlisi, he vist que hi ha diferents casos que demanen un tractament diferent:

- Per una banda, trobem dos dels camps seleccionats on la majoria de valors són "NA". Aquests camps són 'host_acceptance_rate' i (sorprenentment) 'square feets'. Optaré per eliminar els dos camps del dataset perquè qualsevol anàlisi basada en ells tindria tant poca significança sobre el conjunt total de les dades que no podria inferir cap informació útil. Per tant, passaré a tenir 30 camps disponibles i hauré de comptar amb el nombre d'habitacions com a indicador de la grandària de la propietat.
- Per l'altre, hi ha camps on hi ha força valors buits, però són camps redundants. En concret, el camp 'neightbourhood'. En aquest cas, un cop he comprovat que el camp 'neightbourhood cleansed' es fiable i està més informat. També puc eliminar-ho. Per tant, passem a 29 camps.
- Finalment, hi ha altres camps que contenen puntualment algun valor "NA" però dintre de paràmetres normals. En aquests casos, miraré si es factible inferir els valors a partir d'altres camps o decidiré el tractament a fer quan faci cada càlcul en funció del propi càlcul i l'efecte que puguin tenir.

```
> sapply(listings, function(x) sum(is.na(x)))

                          id                     host_id
                           0                           0
                   host_name                   host_since
                           0                           0
          host_response_time           host_response_rate
                           0                           0
       host_acceptance_rate         host_listings_count
                           0                           7
      host_identity_verified        host_has_profile_pic
                           0                           0
               neighbourhood      neighbourhood_cleansed
                           0                           0
neighbourhood_group_cleansed                        city
                           0                           0
                     zipcode               property_type
                           0                           0
                   room_type                  accommodates
                           0                           0
                   bathrooms                    bedrooms
                          42                          23
                        beds                   bed_type
                          37                           0
                 square_feet                       price
                       16876                           0
```

En concret, provaré a omplir els NA de bathrooms, bedrooms i beds a partir del k-veïns més propers usant gower amb k=6:

```
> library(VIM)

> newlistings=kNN(listings, k=6)

> nrow(listings)

[1] 17653

> for (x in 1:17653) {if (is.na(listings$beds[x])) {listings$beds[x]<-newlistings[,21
][x]}}

> for (x in 1:17653) {if (is.na(listings$bedrooms[x])) {listings$bedrooms[x]<-newlist
ings[,20][x]}}

> for (x in 1:17653) {if (is.na(listings$bathrooms[x])) {listings$bathrooms[x]<-newli
stings[,19][x]}}
```

Veiem com un cop fet, ja no tenim valors NA en aquests camps.

```
> sapply(listings, function(x) sum(is.na(x)))
                            id                       host_id                     host_name
                             0                             0                             0
                    host_since             host_response_time           host_response_rate
                             0                             0                             0
          host_acceptance_rate           host_listings_count        host_identity_verified
                             0                             7                             0
          host_has_profile_pic                  neighbourhood       neighbourhood_cleansed
                             0                             0                             0
neighbourhood_group_cleansed                           city                       zipcode
                             0                             0                             0
                 property_type                     room_type                  accommodates
                             0                             0                             0
                     bathrooms                     bedrooms                          beds
                             0                             0                             0
                     bed_type                   square_feet                         price
                             0                         16876                             0
                  cleaning_fee               guests_included             availability_365
                             0                             0                             0
            number_of_reviews          review_scores_rating          cancellation_policy
                             0                          3638                             0
             instant_bookable              reviews_per_month                          size
                             0                          3469                             0
```

Només ens queden al camp de m2, que ja hem dit que directament descartarem perquè el nombre d'observacions és insuficient per tractar-ho i a un parell de camps relatius a els comentaris, que no ens interesa omplir-ho perquè son valors que posen els usuaris i que no tenen perquè obeir a normes d'inferència i prefereixo no manipular.

Per acabar, és interesant notar que he decidit crear un camp nou al fitxer a partir de **discretitzar un camp** ja existent. En concret, he afegit un camp "size" que pren valors "small", "mèdium" o "big" en funció dels valors del camp 'bedrooms' per tal de poder fer servir aquesta agrupació a les anàlisis tal i com explico al apartat següent. Ho he fet de la següent manera:

```
> Tam<-cut(listings$bedrooms,breaks=c(0,1,3,10), labels=c('small','medium','big'))

> table(Tam)

Tam
 small medium    big
 11095   5269    781

> listings$size<-Tam

> listings$size[is.na(listings$size)]<-'small'
```

També he hagut de fer algunes transformacions per tal que els tipus dels camps fossin correctes a R. Per exemple:

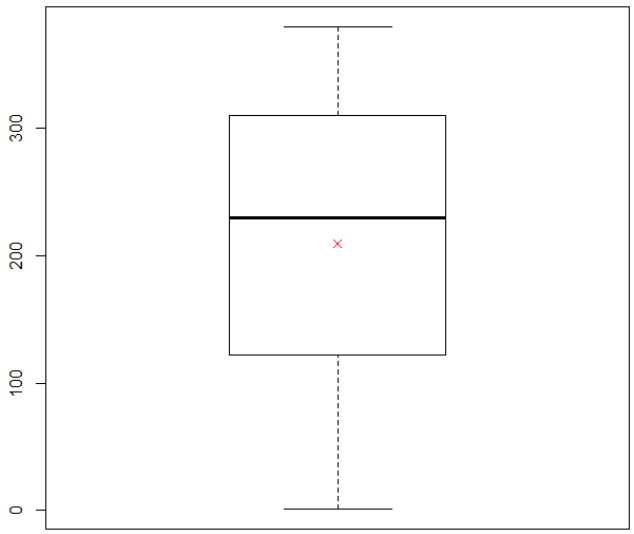
```
> listings$price<-as.numeric(listings$price)
```

Quant al outliers, anem a mirar la variable 'price' per veure com es comporta i si hi ha valors extrems:

Fent un 'fivenum()' veiem que el valor mínim és 1€ i el màxim són 380€ i que sembla que es distribueix força uniformement segons el valor dels quartils.

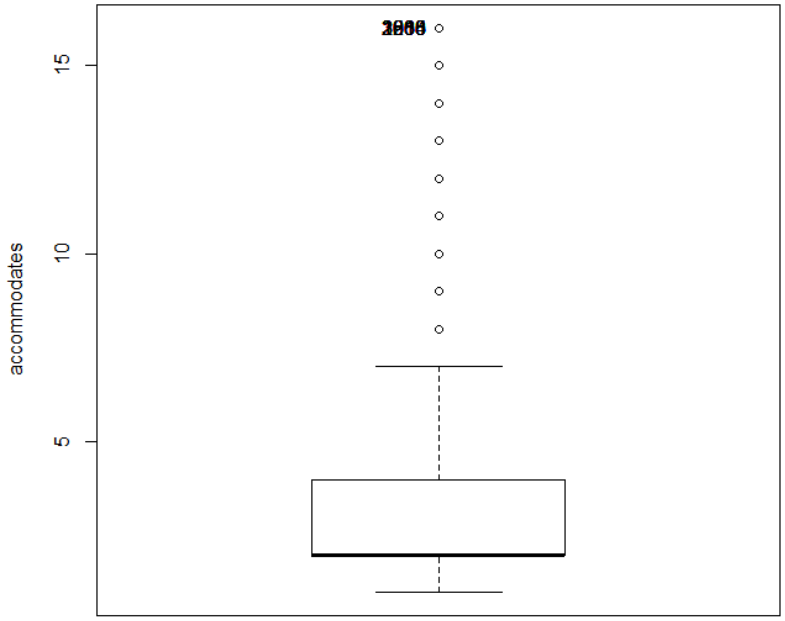
```
> summary(listings$price)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    1.0   122.0   230.0   209.3   310.0   380.0
```

D'entrada no sembla que hi hagi valors extrems però anem a comprovar-ho fent un diagrama de caixes:



Veiem que, efectivament, no trobem valors extrems en aquesta variable. Nota: La creu vermella representa el valor de la mitjana.

Si mirem la variable ‘Accomodates’, però, observem que allà si que trobem valors extrems:



No obstant, si ens fixem en els valors, no sembla pas que aquests valors extrems siguin deguts a errors d’entrada sinó que sembla que corresponen a propietats excepcionalment grans que estan en lloguer. Quadra amb les dades obtingudes de la classificació que hem fet per tamany a partir del nombre d’habitacions, i per tant no em plantejo eliminar-los o corregir-los a base, per exemple, de substituir-los pels valors del 95% o mètodes similars. Ja m’interessa que es mantinguin així per no perdre dades que poden interessar-me posteriorment.

3. Anàlisi de les dades

Abans de res, un resum de com queden les dades després de les primeres transformacions:

```
> summary(listings)
```

```
      id              host_id            host_name          host_since
 Min.   :   10938   Min.   :    10704   Maria  :  187   2016-11-08:  109
 1st Qu.: 4453864   1st Qu.:  5303847   Jordi  :  172   2016-01-12:  108
 Median :10133050   Median : 17782413   Javier :  169   2011-11-10:   77
 Mean   : 9709542   Mean   : 31950899   Claudia:  146   2011-02-17:   65
 3rd Qu.:14861559   3rd Qu.: 49552721   David  :  141   2010-11-26:   62
 Max.   :18110215   Max.   :124704202   Carlos :  121   2009-03-19:   60
                                        (Other):16717   (Other)   :17172
         host_response_time host_response_rate host_acceptance_rate
                  :   7     100%   :10950         :    7
 a few days or more: 494     N/A    : 2082      N/A:17646
 N/A               :2082     99%    :  519
 within a day      :2416     90%    :  382
 within a few hours:3836     97%    :  298
 within an hour    :8818     98%    :  273
                             (Other): 3149
 host_listings_count host_identity_verified host_has_profile_pic
 Min.   :  0.00       :   7                  :    7
 1st Qu.:  1.00      f:7753                 f:   56
 Median :  2.00      t:9893                 t:17590
 Mean   :  9.72
 3rd Qu.:  5.00
 Max.   :207.00
 NA's   :7
                      neighbourhood
                             :6177
 Dreta de l'Eixample         :1304
 El Raval                    : 808
 Vila de GrÃ cia             : 753
 Sants-MontjuÃ¯c             : 691
 La Nova Esquerra de l'Eixample: 666
 (Other)                     :7254
                        neighbourhood_cleansed
 la Dreta de l'Eixample              : 1641
 el Raval                            : 1407
 el Barri GÃ²tic                     : 1161
 la Vila de GrÃ cia                  : 1152
 Sant Pere, Santa Caterina i la Ribera: 1058
 el Poble Sec                        :  989
 (Other)                             :10245
      neighbourhood_group_cleansed                      city
 Eixample            :5752          Barcelona               :17379
 Ciutat Vella        :3910          L'Hospitalet de Llobregat:   61
 Sants-MontjuÃ¯c     :2134          Barcelona               :   50
 Sant MartÃ-         :1881          Barcelona, Catalunya, ES :   36
 GrÃ cia             :1766          barcelona               :   28
 SarriÃ -Sant Gervasi: 768          Ð<U+0091>Ð°Ñ<U+0080>ÑÐµÐ»Ð¾Ð½Ð°       :   22
 (Other)             :1442          (Other)                 :   77
    zipcode              property_type             room_type
```

```
08001  : 1322   Apartment      :15633   Entire home/apt:8869
08003  : 1271   House          :  534   Private room   :8594
08004  : 1179   Bed & Breakfast:  440   Shared room    : 190
08002  : 1120   Loft           :  361
08015  : 1094   Condominium    :  326
08013  :  989   Other          :  156
(Other):10678   (Other)        :  203
 accommodates      bathrooms        bedrooms           beds
Min.   : 1.000   Min.   :0.000   Min.   : 0.000   Min.   : 0.000
1st Qu.: 2.000   1st Qu.:1.000   1st Qu.: 1.000   1st Qu.: 1.000
Median : 2.000   Median :1.000   Median : 1.000   Median : 2.000
Mean   : 3.398   Mean   :1.289   Mean   : 1.529   Mean   : 2.232
3rd Qu.: 4.000   3rd Qu.:1.500   3rd Qu.: 2.000   3rd Qu.: 3.000
Max.   :16.000   Max.   :8.000   Max.   :10.000   Max.   :16.000
                 NA's   :42      NA's   :23       NA's   :37
         bed_type      square_feet         price        cleaning_fee
Airbed       :    8   Min.   :   0.0   Min.   :  1.0   Min.   :  1.00
Couch        :   14   1st Qu.:   0.0   1st Qu.:122.0   1st Qu.:  1.00
Futon        :   63   Median : 129.0   Median :230.0   Median : 26.00
Pull-out Sofa:  151   Mean   : 440.6   Mean   :209.3   Mean   : 34.87
Real Bed     :17417   3rd Qu.: 807.0   3rd Qu.:310.0   3rd Qu.: 68.00
                      Max.   :3444.0   Max.   :380.0   Max.   :113.00
                      NA's   :16876
guests_included  availability_365 number_of_reviews review_scores_rating
Min.   : 1.000   Min.   :  0.0    Min.   :  0.00    Min.   : 20.00
1st Qu.: 1.000   1st Qu.: 73.0    1st Qu.:  1.00    1st Qu.: 86.00
Median : 1.000   Median :216.0    Median :  7.00    Median : 92.00
Mean   : 1.655   Mean   :191.2    Mean   : 23.27    Mean   : 89.94
3rd Qu.: 2.000   3rd Qu.:304.0    3rd Qu.: 30.00    3rd Qu.: 96.00
Max.   :16.000   Max.   :365.0    Max.   :416.00    Max.   :100.00
                                                    NA's   :3638
     cancellation_policy instant_bookable reviews_per_month
flexible       :4693     f:11902          Min.   : 0.020
moderate       :4491     t: 5751          1st Qu.: 0.440
strict         :8292                      Median : 1.060
super_strict_30: 157                      Mean   : 1.552
super_strict_60:  20                      3rd Qu.: 2.270
                                          Max.   :10.790
                                          NA's   :3469
    size
small :11603
medium: 5269
big   :  781
```

3.1 Selecció dels grups de dades que es volen analitzar/comparar

Bàsicament seleccionaré els següents grups principals a l'hora d'analitzar les dades:

- Tipus d'habitació: En realitat té tres valors depenent de si és un pis sencer, una habitació privada o una habitació compartida.
- Tipus de propietat: Quin tipus d'habitatge genèric és.
- Districte: Faré servir el camp 'neightbourhood_group_cleansed'.
- Capacitat: Faré servir com a indicador el número de 'accomodates'.
- Tamany: Faré servir com a indicador el número d'habitacions. Definiré tres nivells a partir de les dades existents factoritzant el camp de manera que tindré:
    - Petits: Entre 0 i 1 dormitoris
    - Mitjans: 2 i 3 dormitoris
    - Grans: Més de 3 dormitoris

Paral·lelament afegiré la resta de factors per cada anàlisi i en cada cas veure el resultat global i el desglossat per cadascun d'aquests grups (en funció dels que sigui aplicable en cada cas o que tingui més sentit aplicar per tal de respondre la pregunta original que pretenem respondre).

3.2 Comprovació de la normalitat i homogeneïtat de la variància. Si és necessari (i possible) aplicar transformacions que normalitzin les dades.

Podem aplicar algun test de normalitat a diferents camps de la mostra i comprovem que tenen una distribució normal, veient el p-value.

```
> normalityTest(~accommodates, test="ad.test", data=listings)


    Anderson-Darling normality test

data:  accommodates
A = 957.4, p-value < 2.2e-16

> normalityTest(~bedrooms, test="ad.test", data=listings)


    Anderson-Darling normality test

data:  bedrooms
A = 2146.3, p-value < 2.2e-16
```

Quant a aplicar transformacions per normalitzar les dades, un camp que veiem que ho necessita força, a priori, és del de city, on per la mateixa ciutat (Barcelona), trobem valors:

- Barcelona
- barcelona
- Barcelona, Catalunya, ES
- Barcelona[espai]
- Barcelone
- Barcelona[espai]
- Bcn
- Barcellona
- Etc..

```
> levels(listings$city)
```

```
 [1] " Barcelona "
 [2] " Barcelona (Metro L1) Hostafrancs"
 [3] "08014 BARCELONA"
 [4] "å·´å¡<U+009E>ç¾<U+0097>é<U+0082>£"
 [5] "å·´å¡<U+009E>é<U+009A><U+0086>æ<U+008B>¿"
 [6] "Barcellona"
 [7] "barcelona"
 [8] "Barcelona"
 [9] "BARCELONA"
[10] "barcelona "
[11] "Barcelona "
[12] "BARCELONA CLOT"
[13] "Barcelona El RAVAL "
[14] "Barcelona sagrada familia "
[15] "Barcelona Sant Andreu de Palomar"
[16] "Barcelona, Catalunya"
[17] "Barcelona, Catalunya, "
[18] "Barcelona, Catalunya, ES"
[19] "Barcelona, EspaÃ±a"
[20] "Barcelonaneta "
[21] "Barcelone"
[22] "Bcn"
```

...

Faig correccions per tal de homogeneïtzar les dades. Faré servir la llibreria ***Stringr***.

```
> library(stringr)
> listings$city<-str_to_title(listings$city)
> listings$city<-trimws(listings$city)
```

```
> listings$city<-str_replace(listings$city,"Barcelone","Barcelona")

> listings$city<-str_replace(listings$city,"Barcelonaneta","Barcelona")

> listings$city<-str_replace(listings$city,"08014 Barcelona","Barcelona")

> listings$city<-str_replace(listings$city,"Barcellona","Barcelona")

> listings$city<-str_replace(listings$city,"Barcelona, Catalunya, Es","Barcelona")

> listings$city<-str_replace(listings$city,"Barcelona, Catalunya,","Barcelona")

> listings$city<-str_replace(listings$city,"Barcelona, Catalunya","Barcelona")

> listings$city<-str_replace(listings$city,"Barcelona (Metro L1) Hostafrancs","Barcel
ona")

> listings$city<-str_replace(listings$city,"Barcelona Clot","Barcelona")

> listings$city<-str_replace(listings$city,"Barcelona El Raval","Barcelona")

> listings$city<-str_replace(listings$city,"Barcelona Sagrada Familia","Barcelona")

> listings$city<-str_replace(listings$city,"Barcelona Sant Andreu de Palomar","Barcel
ona")

> listings$city<-str_replace(listings$city,"Barcelona, Espa*","Barcelona")

> listings$city<-str_replace(listings$city,"Bcn","Barcelona")

> listings$city<-str_replace(listings$city,"Barcelona (Metro L1) Hostafrancs","Barcel
ona")

> listings$city<-str_replace(listings$city,"Barcelona Sant Andreu De Palomar","Barcel
ona")

> listings$city<-str_replace(listings$city,"Barcelona*","Barcelona")

> listings$city<-str_replace(listings$city,"Barcelona *","Barcelona")

> listings$city<-str_replace(listings$city,"Barcelona(Metro L1) Hostafrancs","Barcelo
na")

> listings$city<-str_replace(listings$city,"Barcelona(Metro L1)  Hostafrancs","Barcel
ona")

> listings$city<-str_replace(listings$city,"Barcelona(Metro L1) Hostafrancs","Barcelo
na")

> listings$city<-str_replace(listings$city,"Barcelona(Metro L1) Hostafrancs","Barcelo
na")

> listings$city<-str_replace(listings$city,"Barcelonaã±A","Barcelona")

> listings$city<-str_replace(listings$city,"Å·´Å¡Žç¼–É‚£","Barcelona")

> listings$city<-str_replace(listings$city,"Å·´Å¡Žéš†Æ‹¿","Barcelona")

> listings$city<-str_replace(listings$city,"Catalunya, Es","Barcelona")

> listings$city<-str_replace(listings$city,"Sant Adriã  De Besã²S","Sant Adria De Bes
os")

> listings$city<-str_replace(listings$city,"Sant Adriã  De Besos","Sant Adria De Beso
s")
```

```
> listings$city<-str_replace(listings$city,"L'hospitalet De Llobregat",
+   "Hospitalet De Llobregat")

> listings$city<-str_replace(listings$city,"Sagrada Familia, Barcelona","Barcelona")

> listings$city<-str_replace(listings$city,"Zona Forum","Barcelona")

> listings$city<-str_replace(listings$city,"Zona Franca","Barcelona")

> listings$city<-str_replace(listings$city,"Can Tries","Barcelona")

> listings$city<-str_replace(listings$city,"Corcega","Barcelona")

> listings$city<-as.factor(listings$city)
```

Veiem com ha quedat finalment els valors del camp city un cop correctament normalitzats:

```
> levels(listings$city)

 [1] "Barcelona"
 [2] "Barcelona(Metro L1) Hostafrancs"
 [3] "Caldes De Montbui"
 [4] "Ð<U+0091>ð°Ñ<U+0080>ÑÐµð»Ð¾Ð½Ð°"
 [5] "Ð<U+0091>ð°Ñ<U+0080>ÑÐµð»Ð¾Ð½Ð°, Ð<U+0098>ÑÐ¿Ð°Ð¸Ñ"
 [6] "Eixample (L')"
 [7] "Hospitalet De Llobregat"
 [8] "Montcada Y Reixac"
 [9] "Sant Adria De Besos"
[10] "Santa Coloma De Gramenet"
[11] "St Cugat Del Vallã¨S"
[12] "Tordera"
```

De la mateixa manera, faig correccions anàlogues al camp 'neighbourhood_group_cleansed'. En aquest camp encara és més important perquè és un dels camps que utilitzarem per agrupar les dades.

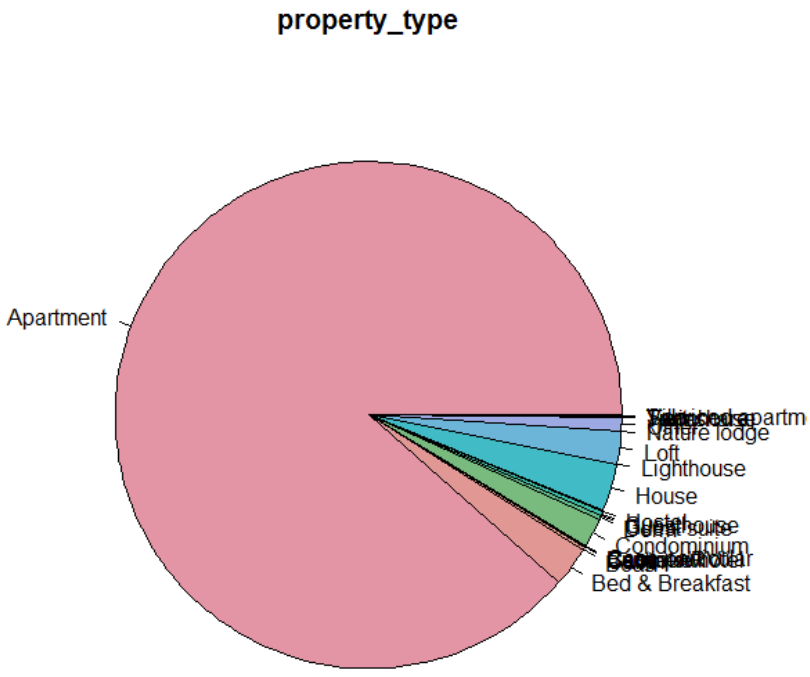
```
> levels(listings$neighbourhood_group_cleansed)

 [1] "Ciutat Vella"         "Eixample"             "GrÃ cia"
 [4] "Horta-GuinardÃ³"      "Les Corts"            "Nou Barris"
 [7] "Sant Andreu"          "Sant MartÃ-"          "Sants-MontjuÃ¨c"
[10] "SarriÃ -Sant Gervasi"

> listings$neighbourhood_group_cleansed<-str_replace(listings$neighbourhood_group_cleansed,"GrÃ cia",
+   "Gràcia")

> listings$neighbourhood_group_cleansed<-str_replace(listings$neighbourhood_group_cleansed,
+   "Horta-GuinardÃ³","Horta-Guinardó")

> listings$neighbourhood_group_cleansed<-str_replace(listings$neighbourhood_group_cleansed,"Sant MartÃ-",
+   "Sant Martí")
```

```
> listings$neighbourhood_group_cleansed<-str_replace(listings$neighbourhood_group_cle
ansed,
+   "Sants-MontjuÃ¯c","Sants-Montjuíc")

> listings$neighbourhood_group_cleansed<-str_replace(listings$neighbourhood_group_cle
ansed,
+   "SarriÃ -Sant Gervasi","Sarrià-Sant Gervasi")

> listings$neighbourhood_group_cleansed<-as.factor(listings$neighbourhood_group_clean
sed)

> levels(listings$neighbourhood_group_cleansed)

 [1] "Ciutat Vella"        "Eixample"            "Gràcia"
 [4] "Horta-Guinardó"      "Les Corts"           "Nou Barris"
 [7] "Sant Andreu"         "Sant Martí"          "Sants-Montjuíc"
[10] "Sarrià-Sant Gervasi"
```

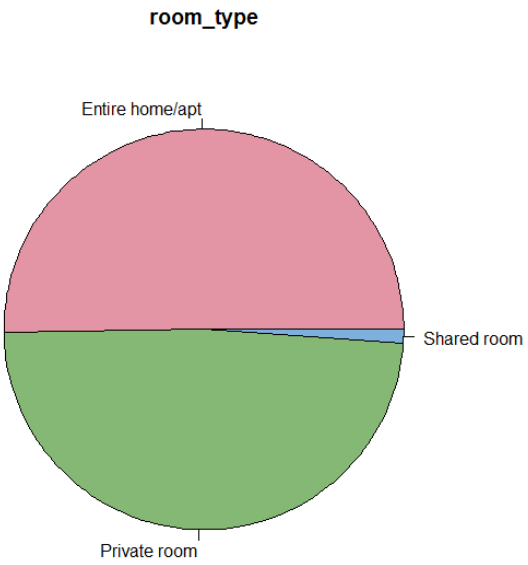
3.3. Aplicació de proves estadístiques per comparar els grups de dades.

El primer que podem fer es veure gràficament com es distribueixen els grups.

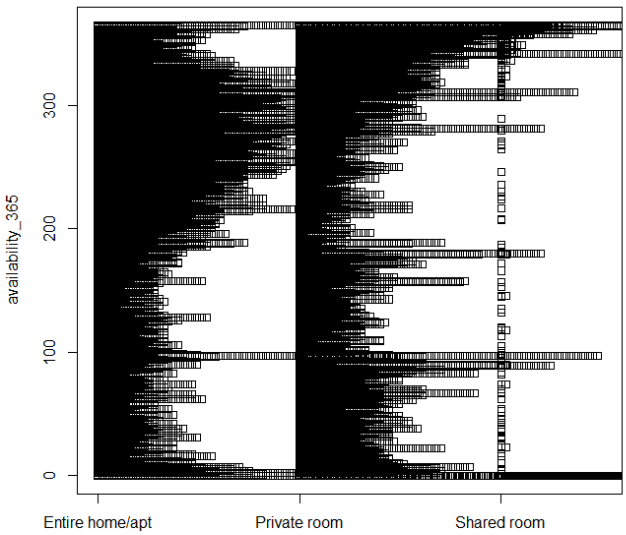
Quant al tipus de propietat, veiem que la majoria corresponen a la categoria "Apartament".



Quant al tipus d'habitació, veiem que les dades pràcticament es reparteixen al 50% entre lloguer d'apartaments complets i lloguer d'habitacions privades, deixant un petit percentatge d'habitacions compartides.



Podem mirar si el fet de tenir poca oferta es degut a que no hi ha demanda o a l'inrevés, si la poca oferta fa que la demanda sigui molt gran. Aquest podria ser un indicador de possibilitat de negoci interesant.

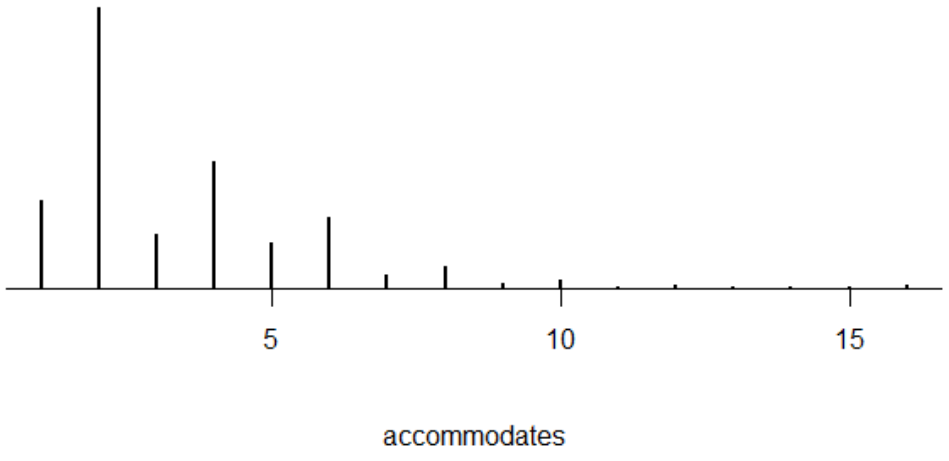


Veiem que segons sembla, **no hi ha oferta perquè tampoc hi ha molta demanda, moltes habitacions compartides no tenen reserva en tot l'any** (la línia del 0 de availability_365).

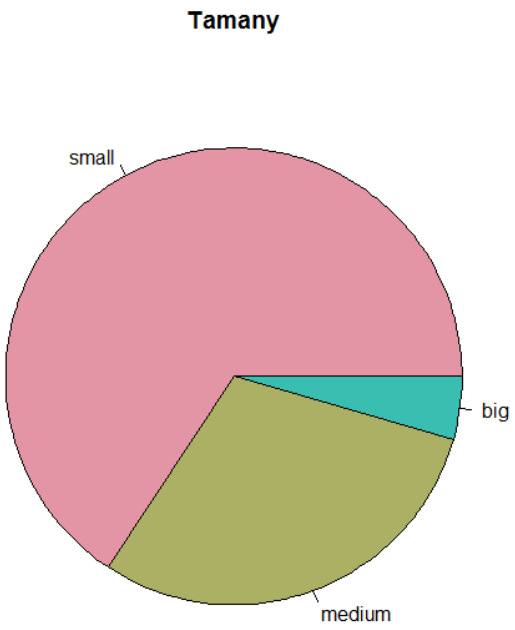
Quant a districtes, podem veure quina és la distribució de freqüències de cadascun i observem que més del 50% de l'oferta es concentra entre Ciutat Vella i l'Eixample.

```
counts:
neighbourhood_group_cleansed
       Ciutat Vella            Eixample              Gràcia      Horta-Guinardó           Les Corts
               3910                5752                1766                 574                 364
         Nou Barris         Sant Andreu          Sant Martí      Sants-Montjuïc Sarrià-Sant Gervasi
                209                 295                1881                2134                 768

percentages:
neighbourhood_group_cleansed
       Ciutat Vella            Eixample              Gràcia      Horta-Guinardó           Les Corts
              22.15               32.58               10.00                3.25                2.06
         Nou Barris         Sant Andreu          Sant Martí      Sants-Montjuïc Sarrià-Sant Gervasi
               1.18                1.67               10.66               12.09                4.35
```

Quant a la capacitat oferta, el número més habitual és per dues persones, tot i que hi ha valors significatius fins a 6 persones. A partir d'aquí, tret de 8 persones la majoria d'altres capacitats són més excepcionals.

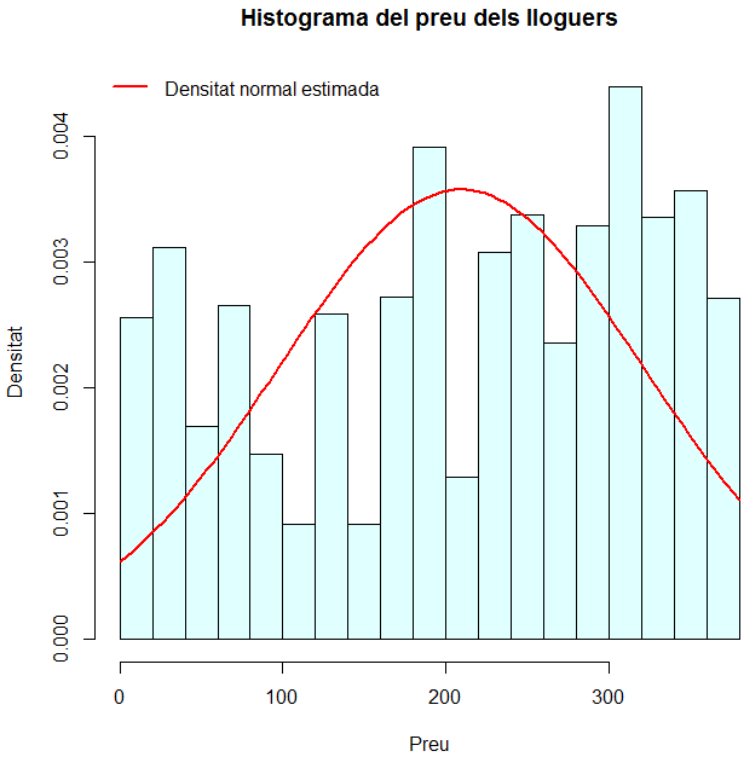


Per acabar aquesta primera ullada a les agrupacions, mirarem com es distribueixen quant al Tamany, camp que hem discretitzat a partir del nombre d'habitacions de la propietat. Veiem que el tamany més usual és el petit que correspon a 1 o 2 habitacions.



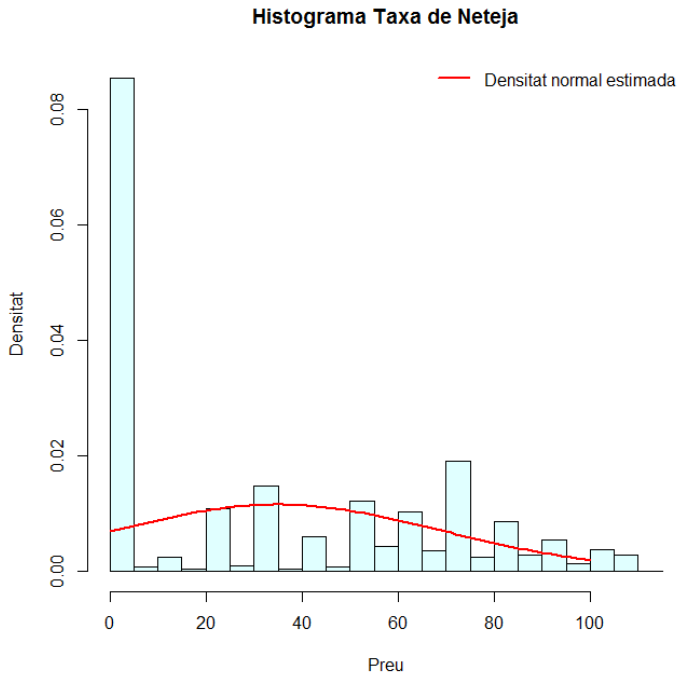
Anem ara ja a mirar una mica la variable de preu, a veure que en podem treure.

Fem un histograma del preu.



Tal i com preveiem al veure els valors dels quartils, sembla que els preus es distribueixen de forma uniforme a tota la franja de preus. Com podem veure, no es correspon a la densitat normal estimada (dibuixada a sobre en vermell).

Quant a l'altre variable de diners, la taxa a pagar en concepte de neteja, en canvi, si que trobem una correspondència més o menys clara (si obviem, es clar, el valor 0 que en aquest cas es molt alt perquè marca clarament la diferencia entre els establiments que han optat per aplicar aquesta taxa i els que no):



Anem ara a fer una anàlisi inferencial sobre el preu: **He trobat per internet que al 2016 la mitjana de preu per nit era de 89€. Vaig a fer un contrast d'hipòtesi per veure si aquest valor es verifica actualment o no**.

<u>Contrast d'hipòtesi</u>

Per tant, la hipòtesi nul.la seria que el preu mitjà fos igual a 89€, mentre que l'alternativa seria que no ho fos. Es a dir, ens trobem davant d'una **hipòtesi bilateral**.

```
> sol.test=t.test(listings$price, mu=99.5, alternative="two.sided", conf.levels=0.95)

> sol.test

        One Sample t-test

data:  listings$price
t = 130.77, df = 17652, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 99.5
95 percent confidence interval:
 207.6462 210.9374
sample estimates:
mean of x
 209.2918
```

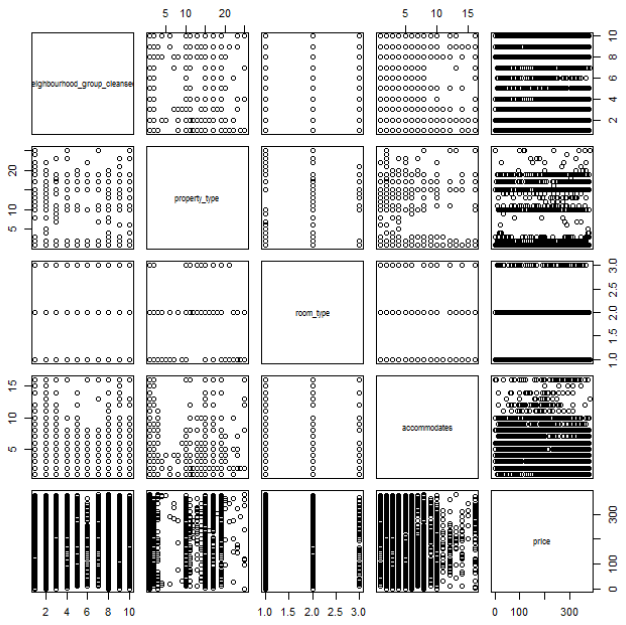
Veiem que el valor de l'estadístic és molt gran i no està prop de zero, per tant, descartem la hipòtesi nul.la. Es a dir, sembla que el preu mig per nit no està ja entorn a aquest valor de 89€. De fet, sembla que està força per sobre.

Influència en el preu

Anem ara veure si podem esbrinar quins factors són més significatius a l'hora de marcar el preu d'un lloguer.

D'entrada, seleccionarem les variables que pensem que poden tenir més relació:
Districte, tipus de propietat, tipus d'habitació, places

Si veiem una primera relació entre elles, no sembla que a priori puguem detectar una dependència clara.



Anem a comprovar-ho aplicant un model de regressió lineal:

```
> resultat<-lm(price ~ neighbourhood_group_cleansed + property_type + room_type + acc
ommodates,
+   data=preus)

> summary(resultat)
```

```
neighbourhood_group_cleansed[T.Eixample]              -9.123  < 2e-16 ***
neighbourhood_group_cleansed[T.Gràcia]                -4.926 8.47e-07 ***
neighbourhood_group_cleansed[T.Horta-Guinardó]        -4.083 4.47e-05 ***
neighbourhood_group_cleansed[T.Les Corts]             -4.913 9.06e-07 ***
neighbourhood_group_cleansed[T.Nou Barris]            -6.560 5.52e-11 ***
neighbourhood_group_cleansed[T.Sant Andreu]           -1.334 0.182144
neighbourhood_group_cleansed[T.Sant Martí]            -5.915 3.37e-09 ***
neighbourhood_group_cleansed[T.Sants-Montjuíc]        -3.925 8.71e-05 ***
neighbourhood_group_cleansed[T.Sarrià-Sant Gervasi]   -5.259 1.46e-07 ***

property_type[T.Loft]                                  3.713 0.000206 ***
property_type[T.Nature lodge]                          0.768 0.442477
property_type[T.Other]                                 0.264 0.791629
property_type[T.Serviced apartment]                    1.425 0.154303
property_type[T.Tent]                                  1.256 0.208984
property_type[T.Timeshare]                            -0.029 0.977013
property_type[T.Townhouse]                             0.080 0.935852
property_type[T.Train]                                 0.521 0.602229
property_type[T.Villa]                                 1.054 0.292038
room_type[T.Private room]                              0.661 0.508899
room_type[T.Shared room]                              -4.009 6.12e-05 ***
accommodates                                         -14.682  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Si ens fixem en el resultat, mirant el nivell de significança veiem que el que sembla influir més al preu és el Districte, les places i el fet de ser habitació compartida. A priori, els resultats semblen prou lògics. El tipus de propietat no sembla influir excepte en el cas dels Lofts (molt probablement degut a la poca disponibilitat d'aquest tipus d'habitatge).

Per tant, podem refer el model només amb aquestes variables:

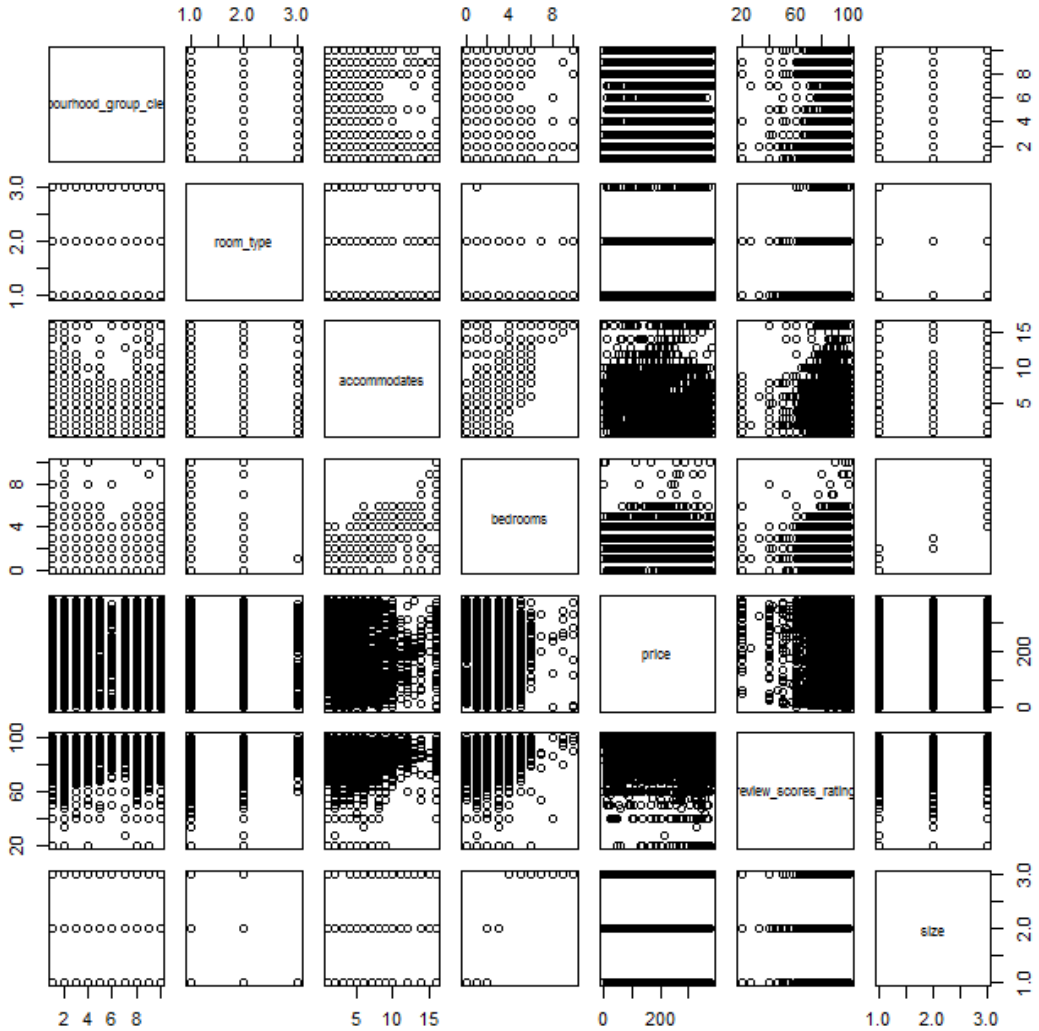
```
> preus<-listings[,c(13,17,18,24)]

> resultat<-lm(price ~ neighbourhood_group_cleansed + room_type + accommodates, data=
preus)
```

Veiem que el model es manté més o menys igual traient la variable, tot i que empitjora lleugerament perquè el valor de R-quadrat ajustat a disminuït una mica.

```
Residual standard error: 109.8 on 17640 degrees of freedom
Multiple R-squared:  0.03113,   Adjusted R-squared:  0.03047
F-statistic: 47.23 on 12 and 17640 DF,  p-value: < 2.2e-16
```

Provem a introduir altres variables a veure si afecten: ***número d'habitacions, tamany, i rati de comentaris***.

```
> preus<-listings[,c(13,17,18,20,24,29,33)]
```



Veiem que hem millorat el model, incrementant el valor R-quadrat ajustat a 0.04924 i reduint lleugerament l'error residual. A més, comprovem que el valor de significança de les noves variables és bo.

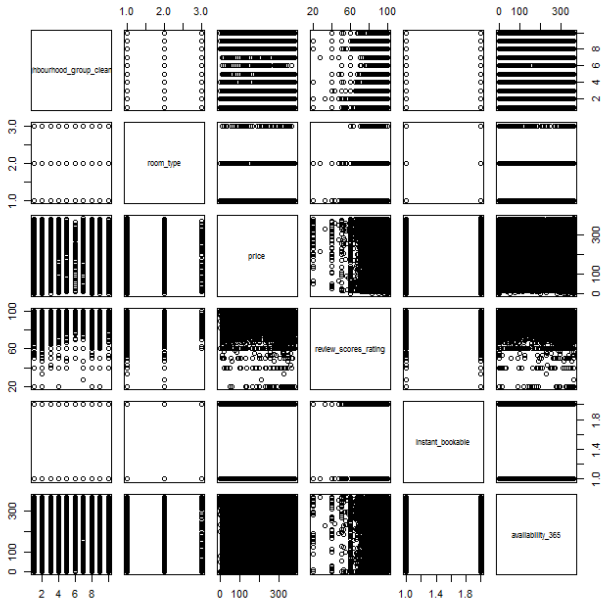
```
> resultat<-lm(price ~ neighbourhood_group_cleansed + room_type + acco
mmodates + bedrooms +
+    review_scores_rating + size, data=preus)

neighbourhood_group_cleansed[T.Eixample]            -7.230 5.09e-13 ***
neighbourhood_group_cleansed[T.Gràcia]              -4.051 5.14e-05 ***
```

```
neighbourhood_group_cleansed[T.Horta-Guinardó]       -3.186 0.001445 **
neighbourhood_group_cleansed[T.Les Corts]            -3.392 0.000696 ***
neighbourhood_group_cleansed[T.Nou Barris]           -5.776 7.83e-09 ***
neighbourhood_group_cleansed[T.Sant Andreu]          -0.528 0.597530
neighbourhood_group_cleansed[T.Sant Martí]           -4.963 7.02e-07 ***
neighbourhood_group_cleansed[T.Sants-Montjuíc]       -2.253 0.024261 *
neighbourhood_group_cleansed[T.Sarrià-Sant Gervasi]  -3.250 0.001155 **
room_type[T.Private room]                            -5.130 2.94e-07 ***
room_type[T.Shared room]                             -8.516  < 2e-16 ***
accommodates                                         -0.631 0.527892
bedrooms                                             -4.732 2.25e-06 ***
review_scores_rating                                 -3.216 0.001302 **
size[T.medium]                                       -6.397 1.64e-10 ***
size[T.big]                                          -1.471 0.141431
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 109.5 on 13998 degrees of freedom
  (3638 observations deleted due to missingness)
Multiple R-squared:  0.05033,   Adjusted R-squared:  0.04924
F-statistic: 46.36 on 16 and 13998 DF,  p-value: < 2.2e-16
```

Influència en la ocupació

Anem a veure ara, anàlogament, quins factors tenen més influència a l'hora d'aconseguir ocupar més dies l'any una propietat.

Inicialment anem a provar amb les variables ***Districte*** novament***, room_type, price*** (però ara no com objectiu sinó com a influència), ***comentaris positius i si es pot reservar instantàniament.***



Si generem un model veiem que totes les variables sembla que tenen significança tret de si es pot reservar al instant.

```
> ocupacio<-listings[,c(13,17,24,29,31,27)]

> resultat2<-lm(availability_365 ~ neighbourhood_group_cleansed + room_type + price +
+   review_scores_rating + instant_bookable, data=ocupacio)

> summary(resultat2)

neighbourhood_group_cleansed[T.Eixample]             10.249  < 2e-16 ***
neighbourhood_group_cleansed[T.Gràcia]                2.793  0.00523 **
neighbourhood_group_cleansed[T.Horta-Guinardó]        6.188 6.25e-10 ***
neighbourhood_group_cleansed[T.Les Corts]             2.452  0.01420 *
neighbourhood_group_cleansed[T.Nou Barris]            5.195 2.07e-07 ***
neighbourhood_group_cleansed[T.Sant Andreu]           0.605  0.54525
neighbourhood_group_cleansed[T.Sant Martí]            8.351  < 2e-16 ***
neighbourhood_group_cleansed[T.Sants-Montjuíc]        3.955 7.70e-05 ***
neighbourhood_group_cleansed[T.Sarrià-Sant Gervasi]   7.980 1.57e-15 ***
room_type[T.Private room]                            -2.261  0.02380 *
room_type[T.Shared room]                              3.280  0.00104 **
price                                                -4.622 3.84e-06 ***
review_scores_rating                                -11.024  < 2e-16 ***
instant_bookable[T.t]                                -0.210  0.83392
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 118 on 14000 degrees of freedom
  (3638 observations deleted due to missingness)
Multiple R-squared:  0.0252,    Adjusted R-squared:  0.02422
F-statistic: 25.85 on 14 and 14000 DF,  p-value: < 2.2e-16
```

Per tant, traiem aquesta i afegirem alguna altre per veure si millorem o empitjorem el resultat. Afegirem la ***política de cancel·lació***, si te ***taxa de neteja,*** i si l'hoste ***te identitat verificada i el nombre de propietats del amfitrió***.

```
> ocupacio<-listings[,c(8,9,13,17,24,25,29,30,27)]

> resultat2<-lm(availability_365 ~ neighbourhood_group_cleansed + room_type + price +
+   review_scores_rating + cancellation_policy + cleaning_fee + host_identity_verifie
d +
+   host_listings_count, data=ocupacio)

> summary(resultat2)

neighbourhood_group_cleansed[T.Eixample]              9.851  < 2e-16 ***
neighbourhood_group_cleansed[T.Gràcia]                3.061 0.002210 **
neighbourhood_group_cleansed[T.Horta-Guinardó]        7.071 1.61e-12 ***
neighbourhood_group_cleansed[T.Les Corts]             2.859 0.004257 **
neighbourhood_group_cleansed[T.Nou Barris]            5.595 2.24e-08 ***
neighbourhood_group_cleansed[T.Sant Andreu]           1.157 0.247168
neighbourhood_group_cleansed[T.Sant Martí]            8.289  < 2e-16 ***
neighbourhood_group_cleansed[T.Sants-Montjuíc]        4.771 1.85e-06 ***
neighbourhood_group_cleansed[T.Sarrià-Sant Gervasi]   8.368  < 2e-16 ***
room_type[T.Private room]                             6.218 5.17e-10 ***
room_type[T.Shared room]                              5.314 1.09e-07 ***
price                                                -3.303 0.000961 ***
review_scores_rating                                 -9.776  < 2e-16 ***
```

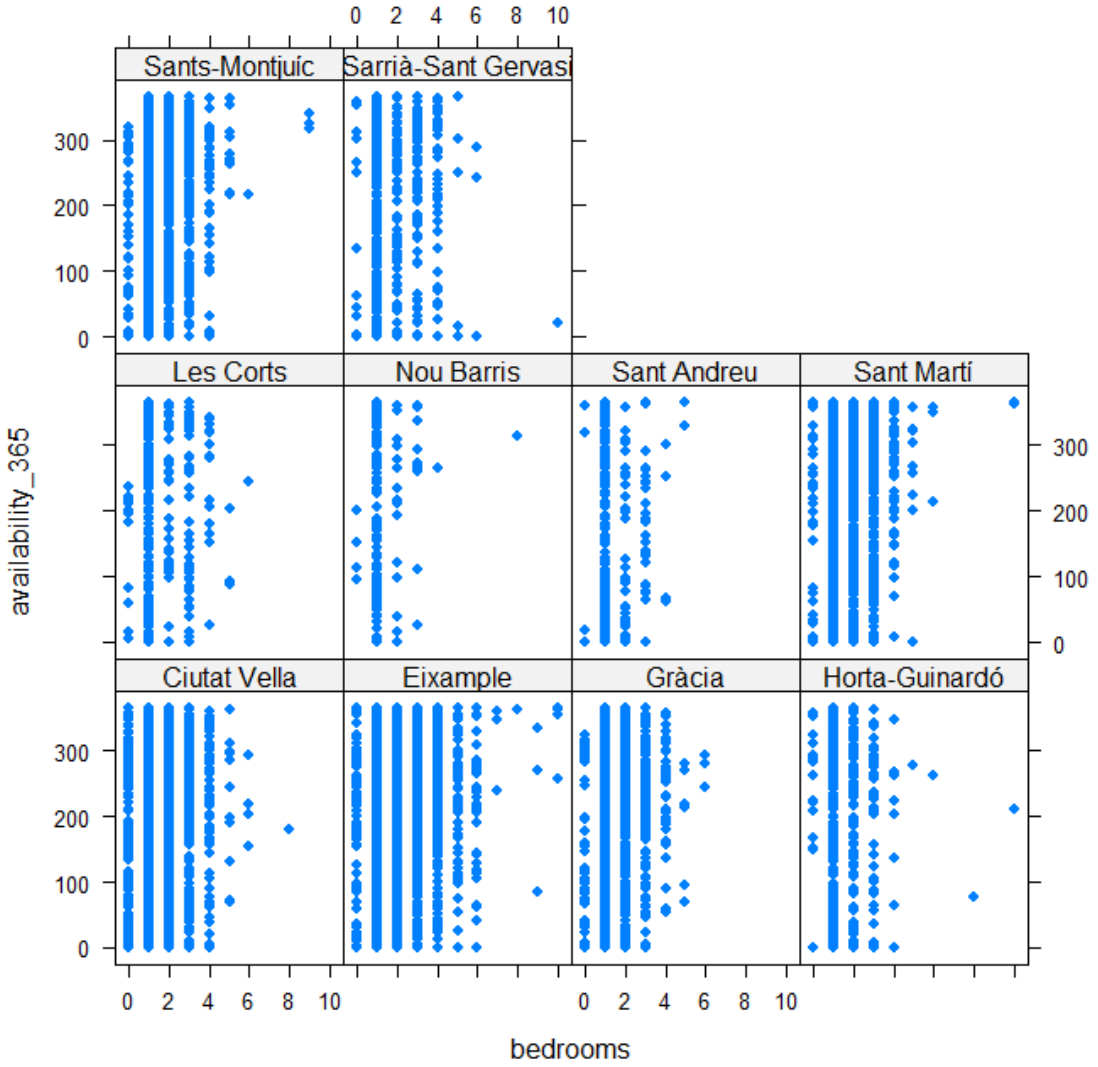
```
cancellation_policy[T.moderate]                          0.490 0.624354
cancellation_policy[T.strict]                            7.593 3.33e-14 ***
cancellation_policy[T.super_strict_30]                   1.777 0.075652 .
cancellation_policy[T.super_strict_60]                  -0.243 0.808071
cleaning_fee                                            11.437  < 2e-16 ***
host_identity_verified[T.t]                             -4.279 1.89e-05 ***
host_listings_count                                      5.184 2.20e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 116.7 on 13993 degrees of freedom
  (3639 observations deleted due to missingness)
Multiple R-squared:  0.04574,   Adjusted R-squared:  0.04437
F-statistic: 33.53 on 20 and 13993 DF,  p-value: < 2.2e-16
```

Veiem que les noves variables tenen totes força significança i hem aconseguit millorar el model fent que el valor e R-quadrat ajustat pugi a 0.044 i disminuint l'error estàndard residual. De les variables introduïdes la menys significativa sembla la política de cancel·lació, però crec que es interessant mantenir-la per la significança que té quan la política és estricte.

4. Representació dels resultats a partir de taules i gràfiques

He anat introduint les gràfiques durant les anàlisis, perquè penso que es més comprensible que no pas fer-ho de forma separada.

Podem afegir un gràfic que penso que es significatiu i que encara no hem fet i que seria el de ocupació en relació a cada Districte i en base al nombre d'habitacions:



Aquest gràfic és important vers a la pregunta inicial perquè ens aclareix com es distribueix la ocupació i ens permet veure que on hi ha més oferta no hi es produeix una saturació de mercat i per tant continua semblant que pot haver lloc per posar més propietats a lloguer.

5. Resolució del problema

A partir dels resultats de l'estudi realitzat, podem concloure com a conclusions principals el següent:

- La majoria d'oferta es concentra en dos Districtes: Ciutat Vella i Eixample
- Tot i així, no sembla que aquesta oferta sobrepassi la demanda, sinó que la demanda es concentra sobretot en aquests dos Districtes
- És important obtenir la confiança del usuari per aconseguir una bona ocupació. Sobretot cuidar els aspectes que hem marcat com significatius per aquest punt.
- Per obtenir un bon preu, ens interesa estar en una bona zona, amb una capacitat del pis inferior o igual a 8 places i amb una puntuació als comentaris alta.

Per tant, de cara a respondre el problema plantejat, tindríem les següents respostes:

- Per un inversor que vulgui adquirir un pis per posar-ho a la venda, ja sap el preu mitjà que pot esperar cobrar per nit i les característiques que ha d'escollir al comprar l'habitatge perquè són les que més influència tenen en el preu que es pot demanar.
- Per una persona que tingui un habitatge i s'estigui plantejant posar-ho en lloguer, podem utilitzar el model generat pel preu per tal de fer una predicció del preu que pot aconseguir i també el model generat per la ocupació per predir quants dies l'any el podria aconseguir llogar.
- Per tots dos tipus de client, tenen clars també els aspectes que com a hoste han de cuidar per tal de generar confiança (i per tant reserves) en les seves propietats.

Com a conclusions addicionals a la pregunta inicial, podem afegir que és molt probable que el malestar dels veïns dels barris de Ciutat Vella i Eixample estigui justificat, havent comprovat que, en efecte, més del 52% de l'oferta d'aquesta plataforma es concentra en aquests dos Districtes.

6. Codi

El codi està adjunt al Github, i correspon al fitxer d'instruccions R utilitzat.