Dades de lloguers habitacionals d'Idealista a les Illes Balears

Carlos A. García

Aconseguir dades

Les dades es corresponen amb les propietats immobiliàries en lloguer a les Illes Balears, segons el principal portal online. Les dades s'han obtingut de la web de gestió immobiliària Idealista mitjançant el seu API d'accés privat. Després s'ha fet un tractament ad-hoc a les dades per tal de centralitzar-les a un únic fitxer que serà el dataset. Pràcticament no s'ha fet cap tractament: només s'han marcat com a NA totes les dades absents. L'accés a les dades, tant sense tractar[1] com tractades[2] és obert.

El conjunt de dades final està compost per 3.600 registres i 28 columnes. Les columnes són:

1. **propertyCode.** Codi de la propietat. Numèrica. Serveix com a clau primària del conjunt de dades.
2. **thumbnail.** URL de la fotografia de presentació de la propietat. Alfanumèrica. No té cap interès per al nostre propòsit.
3. **externalReference.** Referència del llogueter. Alfanumèrica. No té cap interès per al nostre propòsit.
4. **numPhotos.** Nombre de fotografies de l'anunci. Numèrica. S'ha d'estudiar si té correlació amb el preu.
5. **floor.** Pis de la propietat. Categòrica. S'ha d'estudiar si té correlació amb el preu.
6. **price.** Preu per mes. Numèric. Principal variable d'estudi. S'ha de dir que no tots els preus es corresponen amb el valor mensual: hi ha propietats a les que s'indica per una altra unitat de temps (per temporada d'estiu, per exemple). A altres casos, el que s'indica és el preu mínim, que va variant al llarg de l'any. Es planteja com a millora netejar les dades.
7. **propertyType.** Tipus de propietat. Categòrica. S'ha d'estudiar si té correlació amb el preu.
8. **operation.** Règim de l'oferta. Categòrica. En el nostre cas, sempre és lloguer. No té cap interès.
9. **size.** Tamany de la propietat. Numèrica. S'ha d'estudiar si té correlació amb el preu. És molt canviant, ja que el concepte "tamany" és molt flexible. En alguns casos és el tamany de la finca, a altres la superfície construïda, etc.
10. **exterior.** Indicador de si l'habitatge és exterior o no. Booleana. S'ha d'estudiar si té correlació amb el preu.
11. **rooms.** Nombre d'habitacions. Numèrica. S'ha d'estudiar la correlació que té amb el preu. El concepte d'habitació també és flexible depenent de la propietat.
12. **bathrooms.** Nombre de banys. Numèrica. S'ha d'estudiar la correlació que té amb el preu. El concepte de bany també és flexible depenent de la propietat.
13. **address.** Direcció de la propietat. Alfanumèrica. No la considerem.
14. **province.** Província. Categòrica. No té cap interès: sempre són les Illes Balears.
15. **municipality.** Municipi. Alfanumèrica. Té interès tant per mostrar la distribució per municipi com per a avaluar la seva influència al preu. Dir que els municipis no es corresponen totalment amb els municipis reals. Per exemple: Cala Rajada no és municipi, és un poble que pertany a Capdepera. Es planteja com a millora netejar les dades.

[1] https://github.com/caran77/M2.960—Periodisme-de-dades/tree/master/PAC2/data
[2] https://github.com/caran77/M2.960—Periodisme-de-dades/blob/master/PAC2/dataOutput.csv

16. **district.** Districte o barri. Alfanumèrica. En el cas de Palma es poden apreciar diferències notables.
17. **country.** País. Categòrica. No té cap interès al nostre estudi.
18. **latitude.** Coordenada de la propietat. D'especial interès per tal de mostrar les propietats a un mapa.
19. **longitude.** Coordenada de la propietat. D'especial interès per tal de mostrar les propietats a un mapa.
20. **showAddress.** Indicador de si s'ha de mostrar la direcció o no. Booleana. No té cap interès.
21. **url.** Url de la fitxa de la propietat. Alfanumèrica. No té cap interès.
22. **hasVideo.** Indicador de si la fitxa té vídeo associat. Booleana. No té cap interès (el nombre de propietats que tenen vídeo és molt baix)
23. **status.** Estat de l'habitatge. Categòrica. S'ha d'estudiar si té correlació amb el preu.
24. **newDevelopment.** Si és una nova promoció o no. Booleana. S'ha d'estudiar si té correlació amb el preu.
25. **hasLift.** Indicador de si hi ha ascensor. Booleana. S'ha d'estudiar si té correlació amb el preu.

Aplicar mecanismes de selecció i filtratge de les dades

A aquest punt ja tenim les dades a un fitxer CSV extret de un conjunt de fitxers JSON (73 fitxers). El codi que fa aquest tractament està desenvolupat a Python. Eliminem del conjunt de dades les variables que no són del nostre interès; també creem un conjunt de dades específic per a Palma:

```
rentHousesInfo <- read.csv2("dataOutput.csv", header = TRUE, sep = ";", dec = ".")
rentHousesInfo$pricePerMeter <- rentHousesInfo$price / rentHousesInfo$size # Preu per metre
rentHousesInfo <- select(rentHousesInfo, numPhotos, floor, price, size, exterior, rooms,
                         bathrooms, municipality, district, latitude, longitude, status,
                         hasLift, propertyType, pricePerMeter)
# Calculem el rang de preus per a obtenir les gràfiques
rentHousesInfo$priceRang <- floor(rentHousesInfo$price / 1000) * 1000 + 1000
rentHousesInfoPalma <- filter(rentHousesInfo, municipality == 'Palma de Mallorca')
```

Utilitzar descriptors estadístics típics

Mostrem un breu resum dels descriptors estadístics més bàsics pel preu dels habitatges, tant per a les Illes Balears com per a Palma. Mostrem els percentils 25, 50 i 75; la mitjana aritmètica i la desviació típica. Com es pot veure, les dades són molt més canviants al conjunt de la comunitat que a la ciutat.

```
#Percentils, mitjana aritmètica i desviació estàndard del habitatges de les Illes Balears
summary(rentHousesInfo$price)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     350    1000    1450    2085    2300   90000
```

```
sd(rentHousesInfo$price)
```

```
## [1] 2724.9
```

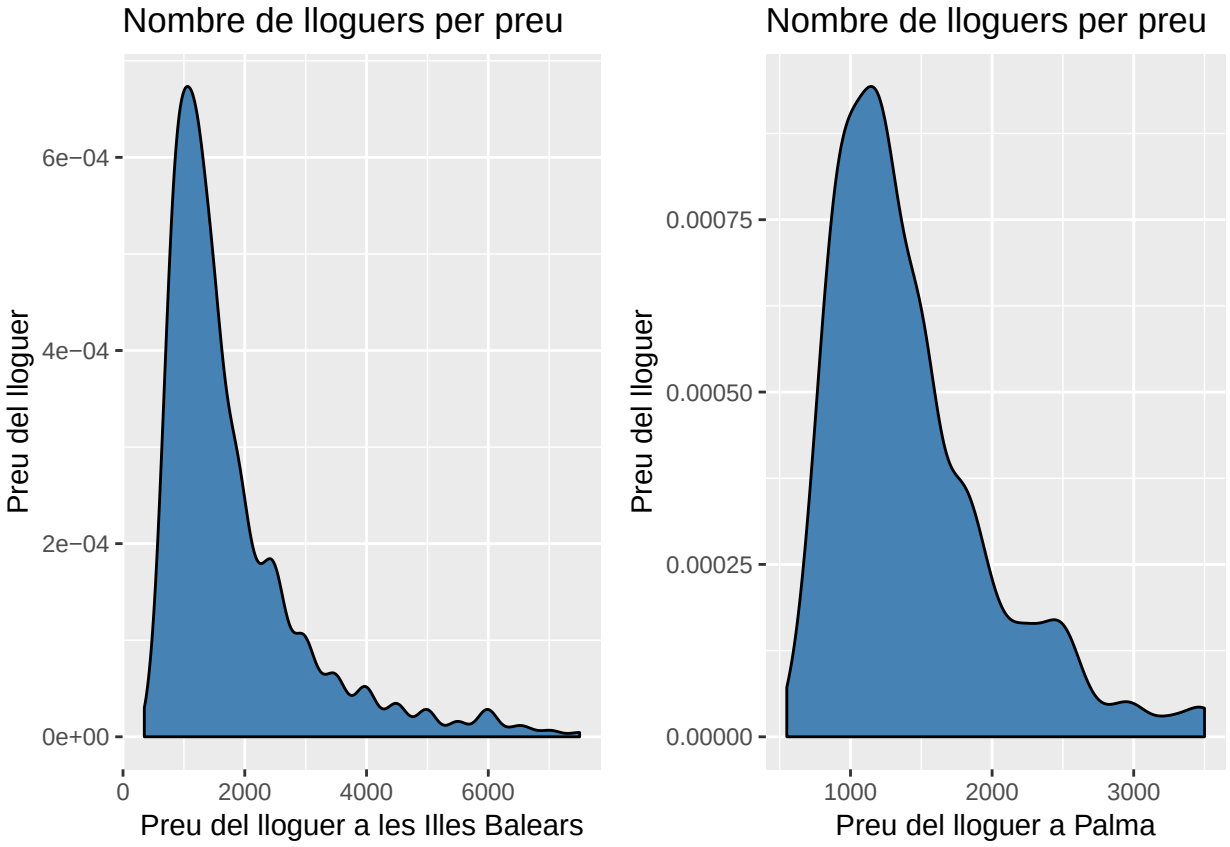
```
#Percentils, mitjana aritmètica i desviació estàndard del habitatges de Palma
summary(rentHousesInfoPalma$price)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     550    1000    1300    1541    1760   15000
```

```r
sd(rentHousesInfoPalma$price)
```

```
## [1] 994.8321
```

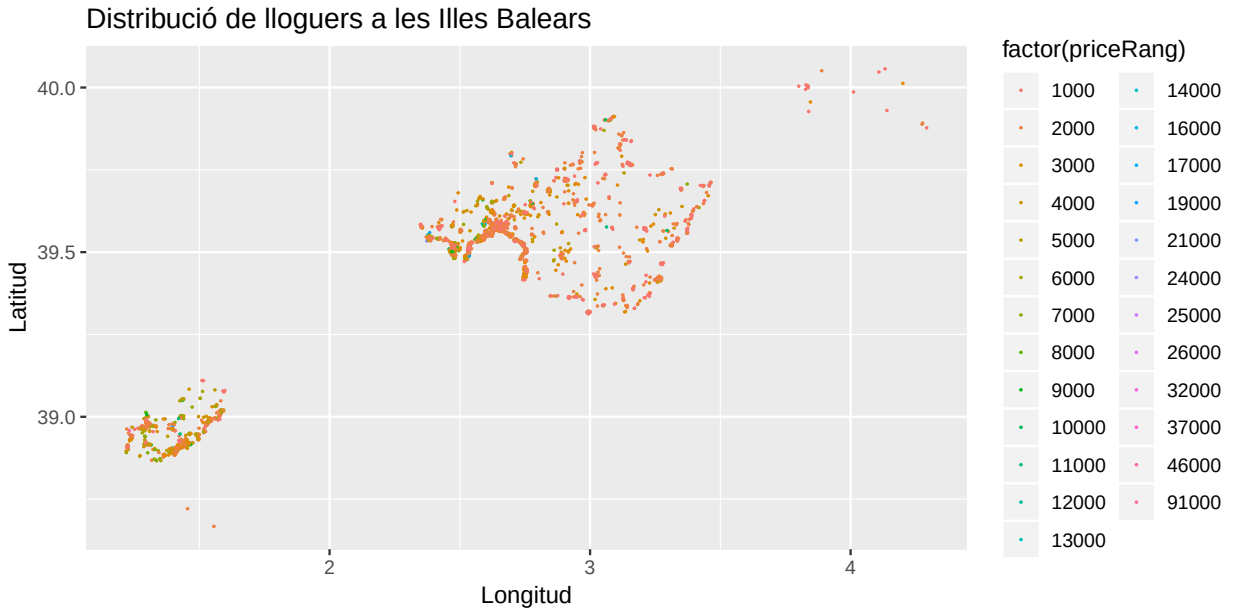
```r
#Mostrem gràficament la dispersió de preus llevant els valors atípics (outliers)
rentHousesInfoPriceSD <- sd(rentHousesInfo$price)
rentHousesInfoPriceMean <- mean(rentHousesInfo$price)
rentHousesInfoWithoutOutliers <- filter(rentHousesInfo, price <= rentHousesInfoPriceMean + 2 * rentHous

rentHousesInfoPriceSDPalma <- sd(rentHousesInfoPalma$price)
rentHousesInfoPriceMeanPalma <- mean(rentHousesInfoPalma$price)
rentHousesInfoWithoutOutliersPalma <- filter(rentHousesInfoPalma, price <= rentHousesInfoPriceMeanPalma

pMallorca <- ggplot() +
  xlab("Preu del lloguer a les Illes Balears") + ylab("Preu del lloguer") + ggtitle("Nombre de lloguers
  geom_density(data=rentHousesInfoWithoutOutliers, aes(x = price), fill="steelblue")
pPalma <- ggplot() +
  xlab("Preu del lloguer a Palma") + ylab("Preu del lloguer") + ggtitle("Nombre de lloguers per preu") +
  geom_density(data=rentHousesInfoWithoutOutliersPalma, aes(x = price), fill="steelblue")
multiplot(pMallorca, pPalma, cols=2)
```



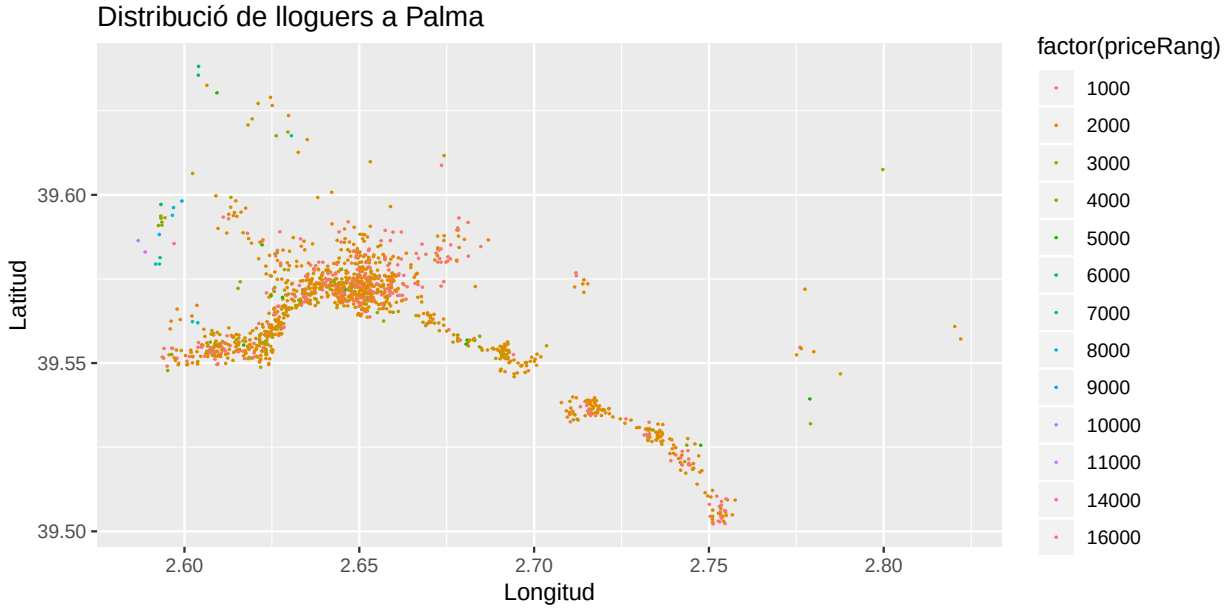
Extreure característiques

Mostrem la distribució de preus, tant per al conjunt de les Illes Balears com per a Palma:

```
ggplot(rentHousesInfo, aes(longitude, latitude)) +
  labs(title="Distribució de lloguers a les Illes Balears", x="Longitud", y = "Latitud") +
  geom_point(size=0.1, aes(colour = factor(priceRang)))
```



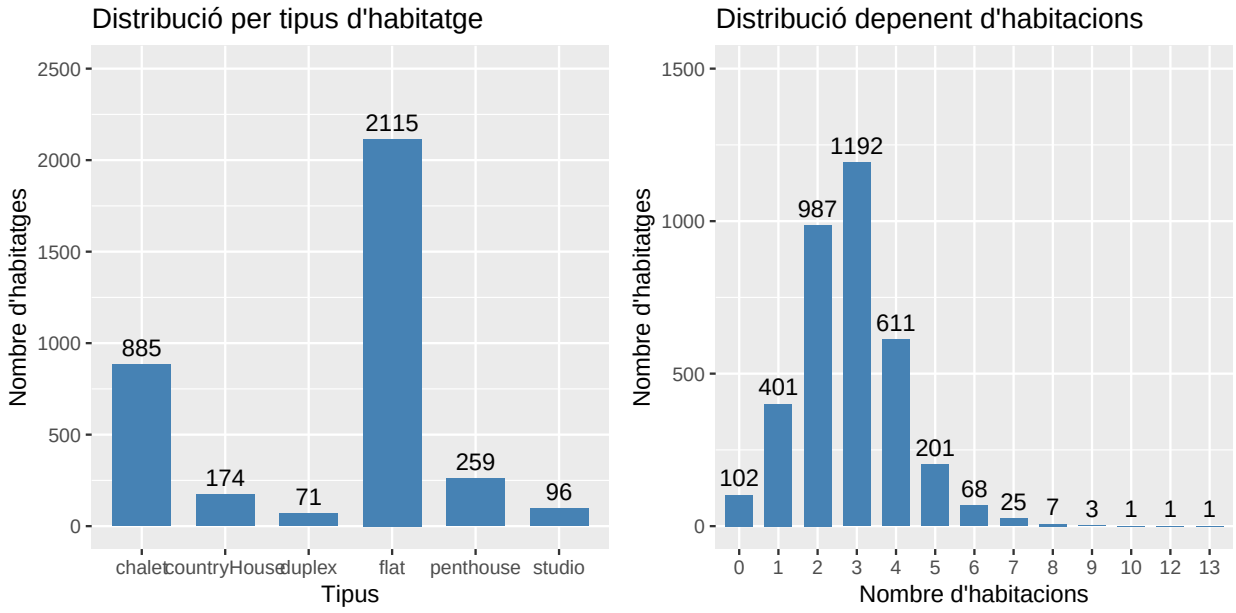

```
ggplot(rentHousesInfoPalma, aes(longitude, latitude)) +
  labs(title="Distribució de lloguers a Palma", x="Longitud", y = "Latitud") +
  geom_point(size=0.1, aes(colour = factor(priceRang)))
```



Es pot veure com els lloguers de Mallorca es concentren a Palma i a la costa; els d'Eivissa a la costa est i sub-oest; pràcticament no existeix ni a Menorca (més enllà de Ciutadella) ni a Formentera (només 2 habitatges). Sembla que els lloguers per Internet funcionen més a les zones urbanes i turístiques. Tot indica que els lloguers a poble funcionen més de forma tradicional.

Mostrem la distribució de lloguers per variable. Tenim la distribució per a totes, mostrem dues com a exemple:

```
pTipus <- ggplot(rentHousesInfo, aes(x=factor(propertyType))) +
  labs(title="Distribució per tipus d'habitatge", x="Tipus", y = "Nombre d'habitatges")+
  geom_bar(stat="count", width=0.7, fill="steelblue") +
  ylim(0, 2500) +
  geom_text(stat="count", aes(label=..count..), vjust=-0.5)

pRooms <- ggplot(rentHousesInfo, aes(x=factor(rooms))) +
  labs(title="Distribució depenent d'habitacions", x="Nombre d'habitacions", y = "Nombre d'habitatges")+
  geom_bar(stat="count", width=0.7, fill="steelblue") +
  ylim(0, 1500) +
  geom_text(stat="count", aes(label=..count..), vjust=-0.5)

multiplot(pTipus, pRooms, cols=2)
```



Trobar el fil argumental a partir de l'anàlisi de les dades

El fil argumental es pot centrar on les dades són més representatives:

1. **A Palma.** Podem tabular el preu per caracterítiques i districte, cercant les principals correlacions.
2. **A la costa, tant de Mallorca com de Eivissa.** Cercar els nuclis on es concentren la major part dels lloguers, establint patrons de preus i característiques.

Com a objectiu interessant es pot afegir la informació de renta per càpita per municipi; d'aquesta forma podrem tenir un indicador del percentatge salarial necessari per a llogar un habitatge a les Illes Balears.