Àlex Penina Aguilera

Tipologia i cicle de vida de les dades

Pràctica 2 (35% nota final)

1) Descripció del dataset. Perquè és important i quina pregunta/problema pretén respondre?

El dataset inclou tots els anuncis d'habitatges en lloguer a la ciutat de Barcelona que hi ha al portal immobiliari d'Habitaclia, el dia 4 de novembre del 2020 (data de l'execució d'extracció de les dades). Per a cada un dels immobles, obtenim les seves principals característiques:

- Barri
- Metres quadrats
- Nombre d'habitacions
- Nombre de lavabos
- Preu
- Tipus d'immoble

He cregut convenient realitzar aquest estudi amb l'objectiu de mostrar la realitat amb la que es troba la població a l'hora de buscar un habitatge de lloguer a Barcelona en un context tan difícil com l'actual, en què la crisi de la Covid-19 ha tingut un impacte notori en tots els àmbits de la societat, incloent l'accés a l'habitatge.

Amb aquestes dades, hi ha la possibilitat de realitzar un estudi sobre el preu mitjà dels habitatges en lloguer anunciats, segmentant-lo per tipus d'habitatge, per barri o per metres quadrats, per exemple.

Així, es podria donar resposta a les preguntes següents, entre d'altres:

- Per quin tipus d'immoble hi ha més anuncis?
- Quin és el preu de lloguer mitjà dels habitatges a Barcelona?
- El barri on es troba l'immoble influeix en el preu del lloguer?
- A quin barri són més cars els lloguers dels immobles? I més barats?
- Quin és l'augment de preu per metre quadrat?
- Quin seria el preu de lloguer estimat d'un pis de $90\mathrm{m}^2$, 3 habitacions i 2 lavabos en el barri de Sant Gervasi?

2) Integració i selecció de les dades d'interès a analitzar.

Les dades fan referència a tots els habitatges en lloguer que hi ha a Barcelona anunciats a la web d'Habitaclia, en data 4 de novembre del 2020.

Les dades s'han recollit utilitzant tècniques de *web scraping* mitjançant codi Python i un fitxer Notebook de JupyterLab, fent ús de diferents llibreries especialitzades, com ara *BeautifulSoup* o *Pandas*.

En el moment de l'extracció, hi havia un total de 13.245 habitatges en lloguer publicats, dividits en 884 pàgines (15 publicacions per pàgina). Degut a això, ha estat necessari crear un bucle que recorregués totes les pàgines i, en cada una d'elles, mitjançant expressions regulars,

s'han extret les variables d'interès i s'han emmagatzemat en una llista. Per últim, s'ha transformat la llista en un *DataFrame* per tal de poder extreure les dades en format CSV.

Finalment, el nostre dataset inclou, per a cada habitatge, les següents variables d'interès a analitzar:

Variable	Tipus	Descripció
Barri	Caràcter	Barri de Barcelona on pertany l'immoble
Tipus_Immoble	Caràcter	Tipologia d'habitatge (pis, casa, xalet, etc.)
Preu	Numèric	Preu mensual exigible per a llogar l'immoble
Metres_Quadrats	Numèric	Nombre de metres quadrats que té l'immoble
Num_Lavabos	Numèric	Nombre de lavabos que té l'immoble
Num_Habitacions	Numèric	Nombre d'habitacions que té l'immoble

3) Neteja de les dades.

3.1) Les dades contenen zeros o elements buits? Com gestionaries aquests casos?

Les dades contenen elements buits, ja que els anuncis no sempre mostren la informació completa dels habitatges. Concretament, existeixen els següents *missings* per a cada variable:

Variable	Nombre de missings
Barri	0
Tipus_Immoble	0
Preu	0
Metres_Quadrats	3
Num_Lavabos	92
Num_Habitacions	3.085

La tècnica aplicada pel tractament de les dades perdudes ha estat la d'ignorar la tupla, és a dir, no fer ús de l'habitatge al que li falten dades.

Tot i que aquesta tècnica pot provocar un biaix en les dades, l'ús d'altres tècniques com la de la mitjana comportaria un biaix encara més gran sobre l'anàlisi de les dades. El nombre d'habitacions, per exemple, està directament lligat amb el nombre de metres quadrats de l'habitatge, fet pel qual no tindria sentit calcular-lo utilitzant la mitjana del nombre d'habitacions de tots els habitatges.

Així doncs, un cop eliminats tots els registres amb elements buits, la nostra base de dades està composta de 10.125 habitatges.

3.2) Identificació i tractament de valors extrems.

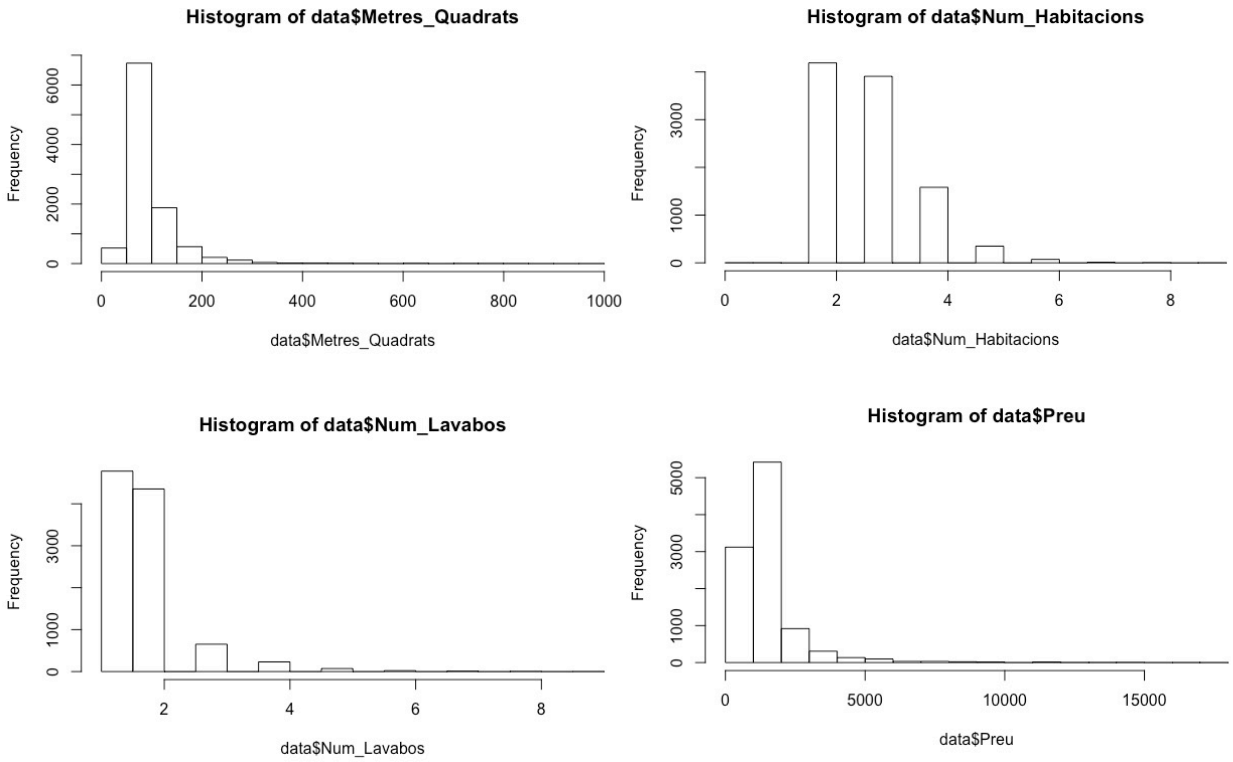
Per identificar els valors extrems utilitzem primer la comanda *summary*, la qual ens mostra els quartils i la mitjana de cada variable numèrica:

```
Metres_Quadrats  Num_Habitacions  Num_Lavabos        Preu
Min.   :  1.00   Min.   :0.00    Min.   :1.000   Min.   :  550
1st Qu.: 68.00   1st Qu.:2.00    1st Qu.:1.000   1st Qu.:  988
Median : 81.00   Median :3.00    Median :2.000   Median : 1200
Mean   : 97.78   Mean   :2.84    Mean   :1.677   Mean   : 1572
3rd Qu.:109.00   3rd Qu.:3.00    3rd Qu.:2.000   3rd Qu.: 1650
Max.   :971.00   Max.   :9.00    Max.   :9.000   Max.   :18000
```

Ja detectem la presència de valors incoherents, com ara habitatges de tan sols $1\mathrm{m}^2$ o amb 0 habitacions. També s'observen valors extrems, com habitatges de $971\mathrm{m}^2$, amb 9 lavabos o amb un preu de 18.000€.

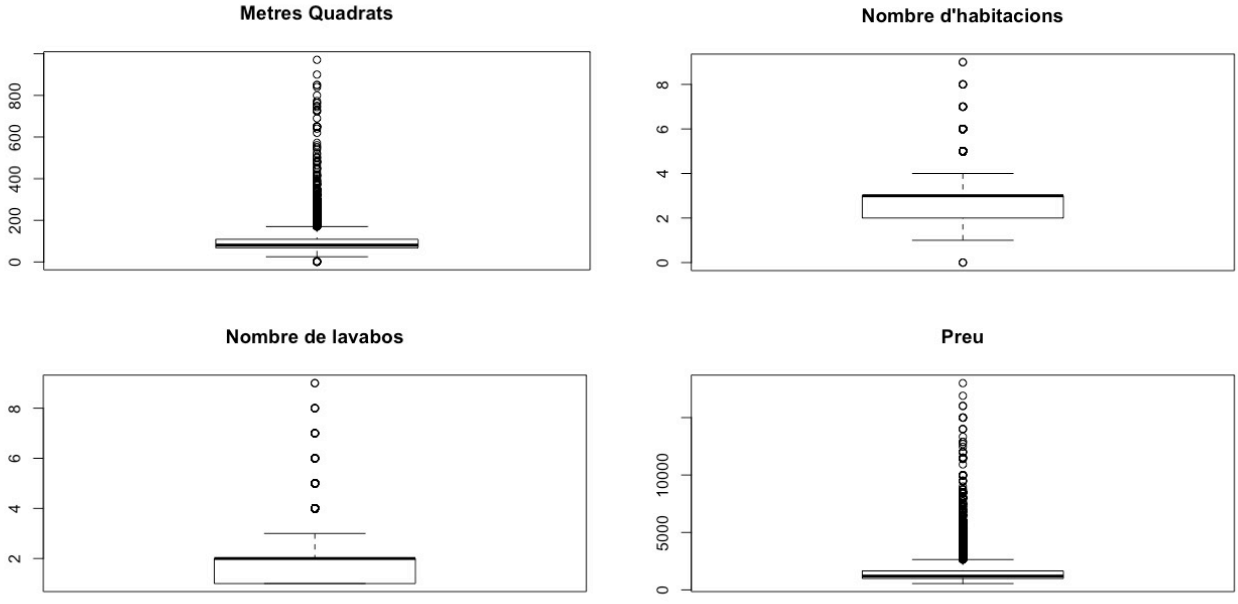
Seguim detectant valors extrems amb l'ajuda de gràfics:

- Histogrames:



Cap de les variables numèriques segueix una distribució aproximadament normal. Totes contenen valors extrems que allarguen l'eix de les X.

- Boxplots:



També observem valors extrems en totes les variables numèriques, els que correspondrien als punts blancs de cada boxplot.

Els valors extrems superiors no ens preocupen, ja que existeixen immobles molt grans i luxosos que prenen valors destacats en totes les variables. El problema el tenim en els punts blancs que apareixen a la part inferior dels boxplots de les variables "Metres quadrats" i "Nombre d'habitacions", que corresponen als valors incoherents que hem detectat en el *summary* anterior.

Així, per exemple, amb la següent instrucció de R:

```
> sum(data$Metres_Quadrats<10)
[1] 9
> sum(data$Num_Habitacions==0)
[1] 3
```

detectem 9 habitatges amb menys de 10m^2 i 3 que no tenen habitacions. Evidentment, es tracta de valors anòmals i, com a tals, els eliminarem de la base de dades.

Així, un cop tractats els *missings* i els *outliers* de la nostra base de dades, aquesta estarà formada finalment per un total de 10.113 habitatges.

4) Anàlisi de les dades i 5) Representació dels resultats a partir de taules i gràfiques.

4.1) Selecció dels grups de dades que es volen analitzar/comparar (planificació dels anàlisis a aplicar).

Si observem el nombre de categories de la variable "Barri":

```
> levels(factor(data$Barri))
 [1] "Badal"                    "Baix Guinardó"
 [3] "Barceloneta"              "Besòs - Maresme"
 [5] "Bon Pastor"               "Camp d´en Grassot - Gràcia N."
 [7] "Camp de l´Arpa"           "Can Baró"
 [9] "Canyelles"                "Carmel"
[11] "Ciutat Meridiana"         "Congrés - Indians"
[13] "Diagonal Mar - La Mar Bella"  "Dreta de l´Eixample"
[15] "El Clot"                  "El Coll"
[17] "Esquerra Alta de l´Eixample"  "Esquerra Baixa de l´Eixample"
[19] "Font d´en Fargas"         "Font de la Guatlla"
[21] "Fort Pienc"               "Glòries El Parc"
[23] "Gòtic"                    "Guinardó"
[25] "Horta"                    "Hostafrancs"
[27] "La Bordeta"               "La Guineueta"
[29] "La Marina-Montjuïc"       "La Marina-Port"
[31] "La Sagrera"               "La Salut"
[33] "La Verneda - La Pau"      "Les Corts"
[35] "Montbau"                  "Navas"
[37] "Pedralbes"                "Poble Sec"
[39] "Poblenou"                 "Porta"
[41] "Prosperitat"              "Provençals del Poblenou"
[43] "Putget - Farró"           "Raval"
```

```
[45] "Roquetes"                    "Sagrada Família"
[47] "Sant Andreu"                 "Sant Antoni"
[49] "Sant Genís dels Agudells"    "Sant Gervasi - Bonanova"
[51] "Sant Gervasi - Galvany"      "Sant Martí"
[53] "Sant Ramon - Maternitat"     "Sants"
[55] "Sarrià"                      "St. Pere - Sta. Caterina - El Born"
[57] "Taxonera"                    "Tres Torres"
[59] "Trinitat Nova"               "Trinitat Vella"
[61] "Turó de la Peira - Can Peguera"  "Vall d´Hebron"
[63] "Vallcarca - Penitents"       "Vallvidrera - Tibidabo - Les Planes"
[65] "Verdun"                      "Vila de Gràcia"
[67] "Vila Olímpica"               "Vilapicina - Torre Llobeta"
```

Obtenim 68 barris diferents, el que ens dificultaria l'anàlisi a l'hora d'aplicar tests estadístics. Per aquesta raó, he decidit reagrupar-los en els 10 districtes de la ciutat de Barcelona, creant una nova variable anomenada "Districte":

```
> levels(factor(data$Districte))
 [1] "Gràcia"          "Sarrià-Sant Gervasi" "Les Corts"       "Sants-Montjuïc"
 [5] "Nou Barris"      "Sant Martí"          "Ciutat Vella"    "Eixample"
 [9] "Sant Andreu"     "Horta-Guinardó"
```

4.2) Comprovació de la normalitat i homogeneïtat de la variància.

El test de Shapiro no es pot realitzar amb el nombre de mostres que tenim, ja que no funciona en mostres més grans de 5.000. No obstant, R inclou altres tests en el paquet "nortest" que permeten analitzar empíricament la normalitat de les variables en mostres grans:

- Test d'Anderson-Darling:

```
> ad.test(data$Metres_Quadrats)

        Anderson-Darling normality test

data:  data$Metres_Quadrats
A = 813.65, p-value < 2.2e-16

> ad.test(data$Num_Habitacions)

        Anderson-Darling normality test

data:  data$Num_Habitacions
A = 732.84, p-value < 2.2e-16
```

```
> ad.test(data$Num_Lavabos)

        Anderson-Darling normality test

data:  data$Num_Lavabos
A = 956.91, p-value < 2.2e-16

> ad.test(data$Preu)

        Anderson-Darling normality test

data:  data$Preu
A = 1165, p-value < 2.2e-16
```

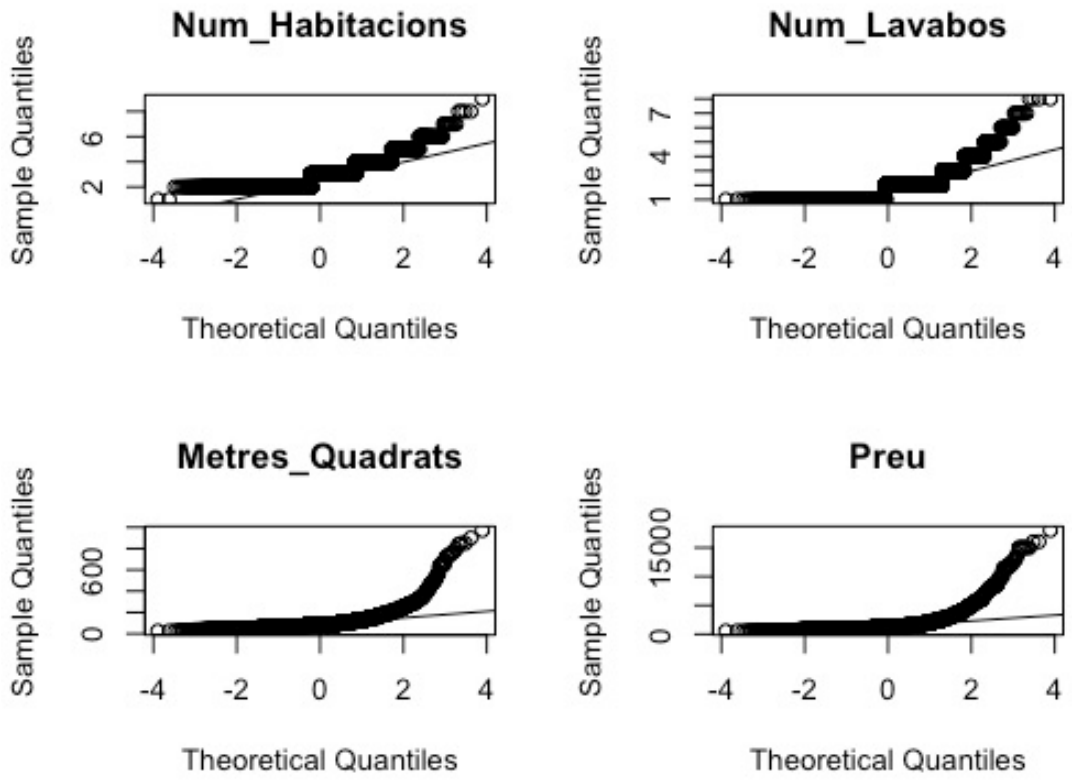
- Test Lilliefors (Kolmogorov-Smirnov):

```
> lillie.test(data$Metres_Quadrats)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  data$Metres_Quadrats
D = 0.20138, p-value < 2.2e-16

> lillie.test(data$Num_Habitacions)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  data$Num_Habitacions
D = 0.2434, p-value < 2.2e-16
```

```
> lillie.test(data$Num_Lavabos)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  data$Num_Lavabos
D = 0.26979, p-value < 2.2e-16

> lillie.test(data$Preu)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  data$Preu
D = 0.23541, p-value < 2.2e-16
```

En ambdues proves ratifiquem que les variables numèriques del nostre dataset no segueixen una distribució normal, ja que el p-valor és molt inferior a 0.05 (nivell de significació del 5%) i, per tant, rebutgem la hipòtesi nul·la de normalitat en les dades.

Ho comprovem gràficament:



Les dades no segueixen la línia de distribució ajustada i, per tant, corroborem la no normalitat en les dades.

Per a estudiar l'homogeneïtat de la variància tindrem en compte els grups definits a partir de la nova variable "Districte". Aplicarem el test de Levene, el qual avalua si les variàncies de les dades en els diferents districtes són iguals o no:

```
> for (i in seq_along(quanti)) print(leveneTest(data[,quanti[i]], data$Districte))
Levene's Test for Homogeneity of Variance (center = median)
         Df F value    Pr(>F)
group     9  44.862 < 2.2e-16 ***
      10103
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Levene's Test for Homogeneity of Variance (center = median)
         Df F value    Pr(>F)
group     9  41.084 < 2.2e-16 ***
      10103
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Levene's Test for Homogeneity of Variance (center = median)
         Df F value    Pr(>F)
group     9   129.9 < 2.2e-16 ***
      10103
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Levene's Test for Homogeneity of Variance (center = median)
         Df F value    Pr(>F)
group     9  79.562 < 2.2e-16 ***
      10103
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Obtenim un p-valor molt inferior al 0.05 en totes les variables i, per tant, amb un nivell de significació del 5%, rebutgem la hipòtesi nul·la d'homogeneïtat de les variàncies entre els grups i podem afirmar que existeixen diferències significatives entre les variàncies de les dades en els diferents districtes de Barcelona.

4.3) Aplicació de proves estadístiques per comparar els grups de dades. En funció de les dades i de l'objectiu de l'estudi, aplicar proves de contrast d'hipòtesis, correlacions, regressions, etc. Aplicar almenys tres mètodes d'anàlisi diferents.

Anem a donar resposta a les preguntes que ens hem formulat a l'inici d'aquest estudi, a partir de varis mètodes d'anàlisi:

- Per quin tipus d'immoble hi ha més anuncis?

Fem un CrossTable de la variable "Tipus_Immoble" per a que ens mostri les freqüències absolutes i relatives:

```
   Cell Contents
|-------------------------|
|                       N |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  10113
```

Apartamento	Ático	Casa	Chalet	Dúplex
1063	397	94	5	135
0.105	0.039	0.009	0.000	0.013

Estudio	Loft	Piso	Planta	Torre
13	23	8335	42	2
0.001	0.002	0.824	0.004	0.000

Tríplex
4
0.000

Els tipus d'immoble pels quals es publiquen més anuncis són els Pisos (82,4%), seguit dels Apartaments (10,5%) i dels Àtics (3,9%).

- Quin és el preu de lloguer mitjà dels habitatges a Barcelona?

Amb la senzilla funció mean() de R podem donar resposta a aquesta pregunta:

```
> mean(data$Preu)
[1] 1566.114
```

El preu de lloguer mitjà dels habitatges a Barcelona, segons les dades extretes el 4 de novembre del 2020 en el portal d'Habitaclia, és de 1.566,11€/mes.

- El districte on es troba l'immoble influeix en el preu del lloguer?

La pregunta era inicialment formulada per barris, però com els hem agrupat en districtes per a un millor anàlisi, la pregunta també s'ha de modificar en conseqüència.

Donat que tenim més de 2 grups, realitzem el test de Kruskal-Wallis. Es tracta d'un mètode no paramètric per a provar si els diferents grups formen part d'una mateixa població o distribució, segons una determinada variable, que en aquest cas és el Preu.

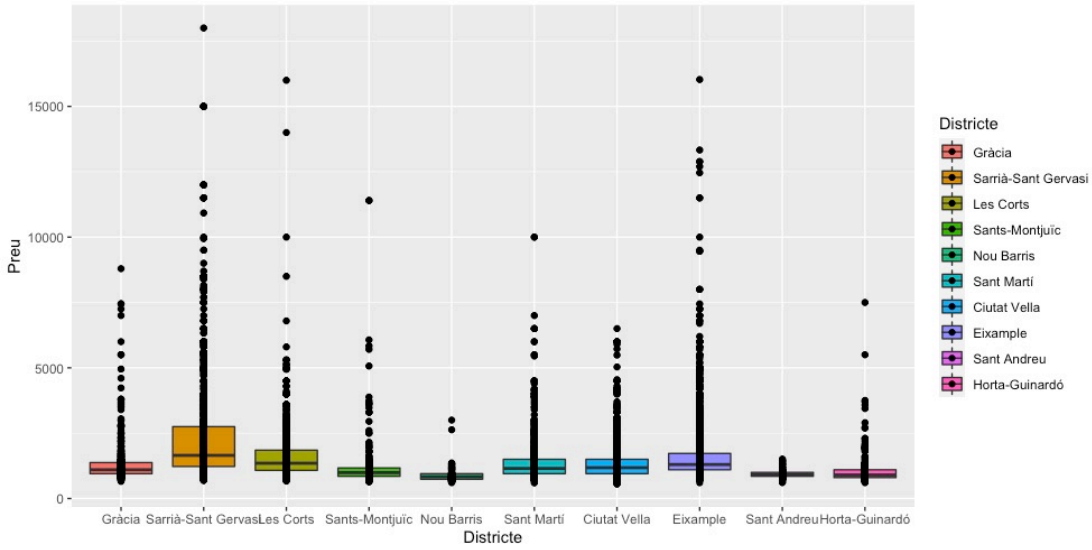
```
> kruskal.test(data$Districte ~ data$Preu)

        Kruskal-Wallis rank sum test

data:  data$Districte by data$Preu
Kruskal-Wallis chi-squared = 1168.6, df = 743, p-value < 2.2e-16
```

Amb un p-valor molt inferior a 0.05, podem rebutjar la hipòtesi nul·la i concloure, amb un nivell de significació del 5%, que els preus dels diferents districtes no formen part d'una mateixa distribució i que, per tant, el districte on es troba l'immoble influeix significativament en el preu del lloguer.

- A quin districte són més cars els lloguers dels immobles? I més barats?

Igual que en la pregunta anterior, modifiquem els barris pels districtes en l'enunciat.

Per a respondre a aquestes preguntes, primer observarem les diferències de preus entre districtes gràficament mitjançant boxplots, gràcies a la llibreria "ggplot2" de R:



A simple vista podem veure com el districte de Sant Gervasi és el que presenta uns preus de lloguer més alts i Nou Barris (o Sant Andreu) els més barats.

Anem a obtenir aquestes diferències empíricament a partir d'un model de regressió lineal, on podrem observar l'efecte que té cada districte sobre el preu. Així, prenent el districte de Gràcia com a referència:

```
> m1 <- lm(Preu ~ Districte, data = data)
> summary(m1)

Call:
lm(formula = Preu ~ Districte, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-1668.7  -498.7  -237.9   101.3 15636.3

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                  1310.65      40.83  32.097  < 2e-16 ***
DistricteSarrià-Sant Gervasi 1053.03      49.98  21.069  < 2e-16 ***
DistricteLes Corts            383.98      60.12   6.387 1.76e-10 ***
DistricteSants-Montjuïc      -205.35      57.88  -3.548 0.000390 ***
DistricteNou Barris          -428.92     102.40  -4.189 2.83e-05 ***
DistricteSant Martí           113.67      55.92   2.033 0.042123 *
DistricteCiutat Vella          47.20      50.37   0.937 0.348701
DistricteEixample             288.02      45.72   6.300 3.11e-10 ***
DistricteSant Andreu         -367.21      85.91  -4.275 1.93e-05 ***
DistricteHorta-Guinardó      -259.89      70.97  -3.662 0.000251 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1146 on 10103 degrees of freedom
Multiple R-squared:  0.1066,    Adjusted R-squared:  0.1058
F-statistic:   134 on 9 and 10103 DF,  p-value: < 2.2e-16
```

Observem com tots els districtes, excepte Ciutat Vella, presenten preus significativament diferents que els de Gràcia (p-valor < 0.05). El més car és Sarrià-Sant Gervasi (1053€ més de mitjana que a Gràcia) i el més barat és Nou Barris (429€ menys de mitjana que a Gràcia).

Tot i això, podem observar com la variable Districte només explica el 10,58% de la variabilitat que hi ha en els preus (Adjusted R-squared).

- Quin és l’augment de preu per metre quadrat?

Realitzem un altre model de regressió lineal per observar en quina mesura augmenta el preu de lloguer per cada metre quadrat:

```
> m2 <- lm(Preu ~ Metres_Quadrats, data = data)
> summary(m2)

Call:
lm(formula = Preu ~ Metres_Quadrats, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-8347.3  -307.6  -126.7   117.9 14660.8

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     50.0927    15.5495   3.221  0.00128 **
Metres_Quadrats 15.5068     0.1365 113.639  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 803.3 on 10111 degrees of freedom
Multiple R-squared:  0.5609,    Adjusted R-squared:  0.5608
F-statistic: 1.291e+04 on 1 and 10111 DF,  p-value: < 2.2e-16
```

Observem que, de mitjana, cada metre quadrat de l'immoble fa augmentar el preu del lloguer en 15,5€.

Aquest cop obtenim un R-squared del 56,08%, el que indica que la variable "Metres_Quadrats" explica gran part de la variabilitat que hi ha en els preus del lloguer.

- Quin seria el preu de lloguer estimat d'un pis de $90\mathrm{m}^2$, 3 habitacions i 2 lavabos en el districte de Sarrià-Sant Gervasi?

Igual que abans, hem de modificar la pregunta i canviar el barri de Sant Gervasi pel districte de Sarrià-Sant Gervasi.

Per tal de poder fer aquest tipus de prediccions, hem de crear un model de regressió lineal multivariant que inclogui totes aquelles variables que puguin resultar significatives, per tal d'explicar els preus dels pisos amb la màxima precisió possible:

```
> m3 <- lm(Preu ~ Metres_Quadrats+Num_Habitacions+Num_Lavabos+Tipus_Immoble+Districte,
data =data)
> summary(m3)

Call:
lm(formula = Preu ~ Metres_Quadrats + Num_Habitacions + Num_Lavabos +
    Tipus_Immoble + Districte, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-7456.2  -291.7   -57.3   142.3 14146.1

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    641.011     42.713  15.007  < 2e-16 ***
Metres_Quadrats                 13.957      0.231  60.415  < 2e-16 ***
Num_Habitacions               -122.932     11.062 -11.113  < 2e-16 ***
Num_Lavabos                    221.371     14.379  15.395  < 2e-16 ***
Tipus_ImmobleÁtico            -152.481     44.993  -3.389 0.000704 ***
Tipus_ImmobleCasa             -604.618     92.299  -6.551 6.01e-11 ***
Tipus_ImmobleChalet            955.289    348.242   2.743 0.006096 **
Tipus_ImmobleDúplex           -448.237     69.955  -6.408 1.55e-10 ***
Tipus_ImmobleEstudio          -552.287    211.970  -2.605 0.009188 **
Tipus_ImmobleLoft             -799.099    160.334  -4.984 6.33e-07 ***
Tipus_ImmoblePiso             -577.398     24.877 -23.210  < 2e-16 ***
Tipus_ImmoblePlanta           -436.530    119.578  -3.651 0.000263 ***
Tipus_ImmobleTorre           -2365.604    546.141  -4.331 1.50e-05 ***
Tipus_ImmobleTríplex         -1127.519    381.035  -2.959 0.003093 **
DistricteSarrià-Sant Gervasi   134.103     34.361   3.903 9.57e-05 ***
DistricteLes Corts             -57.319     40.224  -1.425 0.154191
DistricteSants-Montjuïc        -46.931     38.410  -1.222 0.221787
DistricteNou Barris           -192.909     67.907  -2.841 0.004509 **
DistricteSant Martí             78.117     37.087   2.106 0.035201 *
DistricteCiutat Vella          -18.491     33.660  -0.549 0.582784
DistricteEixample              104.503     30.484   3.428 0.000610 ***
DistricteSant Andreu          -176.970     56.991  -3.105 0.001906 **
DistricteHorta-Guinardó       -145.950     47.095  -3.099 0.001947 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 759.3 on 10090 degrees of freedom
Multiple R-squared:  0.6085,    Adjusted R-squared:  0.6077
F-statistic: 712.9 on 22 and 10090 DF,  p-value: < 2.2e-16
```

Observem que, amb un nivell de significació del 5%, totes les variables són significativament rellevants per a explicar els preus de lloguer, ja que totes presenten un p-

valor < 0.05. El conjunt d'aquestes variables aconsegueixen explicar el 60,77% de la variabilitat total dels preus.

Així, amb aquest model podem fer la predicció que ens preguntàvem a l'inici amb un nivell de precisió més alt:

```
> predict(m3, newdata = data.frame(Num_Habitacions = 3,
+                                  Num_Lavabos = 2,
+                                  Tipus_Immoble = "Piso",
+                                  Metres_Quadrats = 90,
+                                  Districte = "Sarrià-Sant Gervasi"))
       1
1527.811
```

Segons el model creat, un pis de $90\mathrm{m}^2$, 3 habitacions i 2 lavabos en el districte de Sarrià-Sant Gervasi tindrà un lloguer mensual de 1.527,81€.

6) Resolució del problema. A partir dels resultats obtinguts, quines són les conclusions? Els resultats permeten respondre al problema?

Un cop analitzades les dades extretes del portal web d'Habitaclia, referents als lloguers d'habitatges en data 4 de novembre del 2020, podem concloure que hem donat resposta a totes les preguntes que ens havíem plantejat a l'inici.

D'aquesta manera, ara sabem que el tipus d'immoble que més es publica (i amb molta diferència) són els pisos, que el preu mitjà del lloguer dels habitatges a Barcelona és de 1.566,11€ al mes, o que cada metre quadrat que té l'immoble fa augmentar, de mitjana, en 15,5€ el preu del lloguer.

També hem pogut determinar que el districte on es troba l'immoble influeix significativament en el preu del lloguer i, a partir d'aquí, hem esbrinat que el districte de Barcelona on els lloguers són més cars és el de Sarrià-Sant Gervasi, mentre que els lloguers més barats són a Nou Barris.

Per últim, hem creat un model lineal multivariant a partir de totes les variables de les que disposàvem i hem descobert que totes són significatives per a explicar el preu de lloguer dels immobles. Així, el preu de lloguer d'un habitatge pot variar significativament en funció dels metres quadrats que tingui, del seu nombre d'habitacions i de lavabos, del tipus d'immoble que sigui i del districte on estigui situat. Totes aquestes variables aconsegueixen explicar el 60,77% de la variabilitat que hi ha en els preus de lloguer a la ciutat de Barcelona. D'aquesta manera, a partir d'aquest model podem predir el preu de lloguer d'un habitatge segons les seves característiques, el qual ens pot servir d'ajuda a l'hora de buscar un lloguer que s'ajusti a les nostres necessitats (i a la nostra butxaca).

7) **Codi: Cal adjuntar el codi, preferiblement en R, amb el que s'ha realitzat la neteja, anàlisi i representació de les dades. Si ho preferiu, també podeu treballar en Python.**

S'ha utilitzat R per al tractament de les dades. El codi és el següent:

```
#TIPOLOGIA I CICLE DE VIDA DE LES DADES -- PRAC2

#Lectura de les dades

#Indiquem on es troba el fitxer que volem carregar:
setwd("~/Desktop/Màster UOC Data Science/3r semestre/Tipologia i ci
cle de vida de les dades/PRAC2")
#Carreguem l'arxiu csv:
data<-read.csv2('Habitatges_Lloguer_Barcelona_20201104.csv', sep=",
", fileEncoding = "utf-16")

#Observem de quin tipus és cada variable:
str(data)

#La variable 'Preu' figura com a caràcter. La transformem a numèric
a, eliminant amb la funció gsub() els caràcters no numèrics:
data$Preu <- gsub(" €","", data$Preu)
data$Preu <- gsub(".","", data$Preu, fixed = T)
data$Preu <- as.integer(as.character(data$Preu))
str(data)



#Neteja de les dades:

#Missings:
#Nombre de missings per variable:
sapply(data, function(x) sum(is.na(x)))
#Eliminem els habitatges que continguin algun missing:
data <- na.omit(data)

#Outliers:
#Mostrem els quartils i la mitjana de les variables numèriques:
summary(data)

#Gràfics
#Histogrames
hist(data$Metres_Quadrats)
hist(data$Num_Habitacions)
hist(data$Num_Lavabos)
hist(data$Preu)

#Boxplots
boxplot(data$Metres_Quadrats, main='Metres Quadrats')
boxplot(data$Num_Habitacions, main="Nombre d'habitacions")
boxplot(data$Num_Lavabos, main='Nombre de lavabos')
boxplot(data$Preu, main='Preu')

#Detecció de valors anòmals:
sum(data$Metres_Quadrats<10)
sum(data$Num_Habitacions==0)

#Eliminem els valors anòmals:
data <- data[!(data$Metres_Quadrats<10),]
data <- data[!(data$Num_Habitacions==0),]
```

```r
#Anàlisi de les dades:

#Observem el nombre de categories de la variable Barri:
levels(factor(data$Barri))

#Reagrupem els barris en districtes per reduir el nombre de categor
ies:
install.packages("readxl")
library(readxl)
#Llegim l'Excel que conté les relacions entre els barris i els dist
ricts:
barris <- as.data.frame(read_excel("Barri_Districte.xlsx", sheet =
1))
districtes <- as.data.frame(read_excel("Barri_Districte.xlsx", shee
t = 2))

#Creem la nova variable Districte amb la nova agrupació dels barris
:
for (i in 1:nrow(barris)) data$Districte[as.character(data$Barri) =
= barris[i, 1]] <- barris[i, 2]
data$Districte <- factor(data$Districte, levels = 1:10, districtes$
Districtes)

for (i in 1:nrow(barris)) dat$Barri2[as.character(dat$Barri) == bar
ris[i, 1]] <- barris[i, 2]
dat$Barri2 <- factor(dat$Barri2, levels = 1:10, barris_codi$Barri )

levels(factor(data$Districte))

sapply(data, function(x) sum(is.na(x)))
#Tots els barris han estat assignats al seu corresponent districte.




#Comprovació de la normalitat i homogeneïtat de la variància

#Normalitat
shapiro.test(data$Metres_Quadrats)
#No es pot realitzar el test de Shapiro, ja que la bdd ha de tenir
entre 3 i 5000 registres.

#Utilitzarem altres tests de normalitat, inclosos en el paquet "nor
test":
install.packages("nortest")
library(nortest)

#Prova d'Anderson-Darling:
ad.test(data$Metres_Quadrats)
ad.test(data$Num_Habitacions)
ad.test(data$Num_Lavabos)
ad.test(data$Preu)

#Prova de Lilliefors (Kolmogorov-Smirnov):
lillie.test(data$Metres_Quadrats)
lillie.test(data$Num_Habitacions)
lillie.test(data$Num_Lavabos)
lillie.test(data$Preu)
#Els 2 tests indiquen una NO normalitat de les variables numèriques
.
```

```r
#Gràfics de normalitat:
quanti <- c("Num_Habitacions", "Num_Lavabos", "Metres_Quadrats", "P
reu")
par(mfrow = c(2,2))
for (i in seq_along(quanti)) {
  qqnorm(data[,quanti[i]],main = quanti[i])
  qqline(data[,quanti[i]])
}


#Homogeneïtat de les variàncies:
install.packages("car")
library(car)
for (i in seq_along(quanti)) print(leveneTest(data[,quanti[i]], dat
a$Districte))
#Es rebutja homogeneïtat de les variàncies per a totes les variable
s quantitatives, tenint en compte com a grups els districtes.




#Mètodes d'anàlisi per a donar resposta a les preguntes

#Per quin tipus d'immoble hi ha més anuncis?
#Fem un CrossTable per a donar resposta a aquesta pregunta:
install.packages("gmodels")
library(gmodels)
CrossTable(data$Tipus_Immoble)


#Quin és el preu de lloguer mitjà dels habitatges a Barcelona?
mean(data$Preu)


#El districte on es troba l'immoble influeix en el preu del lloguer
?
#Al tenir més de 2 grups, realizem el test de kruskall-Wallis:
kruskal.test(data$Districte ~ data$Preu)
#Es rebutja la hipòtesi nul·la i, per tant, el districte influeix e
n el preu del lloguer.


#A quin districte són més cars els lloguers dels immobles? I més ba
rats?
#Observem gràficament aquestes diferències mitjançant boxplots dels
preus segons el districte:
install.packages("ggplot2")
library(ggplot2)
ggplot(data, aes(x=Districte, y=Preu, fill=Districte)) + geom_boxpl
ot() + geom_point()

#Fem un model de regressió lineal per a determinar en quina magnitu
d influeix cada districte en el preu del lloguer:
#Ajustem un model de regressió lineal per veure l'efecte de cada ba
rri:
m1 <- lm(Preu ~ Districte, data = data)
summary(m1)


#Quin és l'augment de preu per metre quadrat?
```

```
#Realitzem un altre model de regressió lineal:
m2 <- lm(Preu ~ Metres_Quadrats, data = data)
summary(m2)


#Quin seria el preu de lloguer estimat d'un pis de 90m2, 3 habitaci
ons i 2 lavabos en el barri de Sant Gervasi?
#Creem un model de regressió lineal multivariant per a poder explic
ar amb la màxima precisió possible els preus de lloguer:
m3 <- lm(Preu ~ Metres_Quadrats+Num_Habitacions+Num_Lavabos+Tipus_I
mmoble+Districte, data =data)
summary(m3)
#Totes les variables resulten significatives per al model.

#Amb aquest últim model ja podem fer la predicció:
predict(m3, newdata = data.frame(Num_Habitacions = 3,
                                 Num_Lavabos = 2,
                                 Tipus_Immoble = "Piso",
                                 Metres_Quadrats = 90,
                                 Districte = "Sarrià-Sant Gervasi")
)
#La predicció del preu del pis és de 1.527,81€.

#Per últim, exportem la bdd final en format CSV:
write.csv(data,"~/Desktop/Màster UOC Data Science/3r semestre/Tipol
ogia i cicle de vida de les dades/PRAC2/Habitatges_Lloguer_Barcelon
a_20201104_FINAL.csv")
```

Taula de contribucions del treball.

Com ja li vaig comentar al professor col·laborador de l'assignatura, a l'organitzar-me les tasques del màster per assignatures vaig tardar uns dies de més en assabentar-me que calia formar grups per a la realització de la pràctica i, degut a aquest fet, no vaig trobar cap altre alumne disponible per a formar parella. El professor no va posar-me objecció a que realitzés la pràctica individualment.

Contribucions	Signa
Investigació prèvia	Àlex Penina
Redacció de les respostes	Àlex Penina
Desenvolupament del codi	Àlex Penina