Pràctica 2

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

El dataset és un recull dels últims habitatges en lloguer a Barcelona que s'han publicat a través del portal web habitaclia.com durant el mes d'abril del 2019.
Les dades que s'han extret poden conduir a la realització d'un anàlisi sobre la tendència de la oferta d'habitatges de lloguer a Barcelona i de les seves característiques en relació al preu.

Les dades extretes corresponen al tipus d'habitatge (per exemple si és un pis, apartament o estudi), metres quadrats, número d'habitacions i lavabos, preu de lloguer i una breu descripció personalitzada de l'anunciant.

El conjunt de dades recollit permet analitzar amb més detall quina és la tendència dels habitatges en lloguer i quina variació ha patit en el mes d'abril de l'any 2019. Concretament, les dades del dataset permetran respondre les següents preguntes:

- Quin tipus d'immoble és el mes publicat?
- El preu del m2 es veu afectat per el barri on es troba l'immoble?
- El preu per m2 és veu afectat pel número de metres de l'immoble?
- Quin és l'augment de preu per m2?

Més enllà de les preguntes proposades, si el període temporal del data set fos més llarg també es podria respondre a tipus de preguntes predictives, descriptives, exploratòries o inferencials.

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

Les dades del dataset generat a la pràctica 1 pertanyen als últims habitatges de lloguer publicats al portal de habitaclia.com en el mes d'abril de l'any 2019.

El dataset inclou els següents camps:

Camp	Descripció
Barri	Barri on pertany l'habitatge de lloguer publicat
Habitacions	Número d'habitacions que té l'habitatge
Lavabo	Número de lavabos que té l'habitatge
Metres	Número de metres quadrats que té l'habitatge
Preu	Preu que és demanar per llogar l'habitatge
TipusImmoble	Tipus d'immoble que s'està oferint, que pot ser tipus pis, apartament, dúplex, àtic, casa, loft, estudi, xalet, entre d'altres. És una variable categòrica
Descripció	Breu resum de l'habitatge publicat

3. Neteja de les dades.

Les dades del dataset van ser extretes de la pràctica anterior a partir d'un script generat amb Python i BeautifulSoup. Atès que les dades van ser extrets amb expressions regulars, ens ha simplificat a la tasca a l'hora de realitzar la neteja. Així, la neteja ha consistit en eliminar caràcters especials com el símbol de la moneda de l'Euro a la columna Preu i eliminar el punt que s'utilitzava com a marcador de milers.

```
dat$Preu <- gsub(" â□¬","", dat$Preu)
dat$Preu <- gsub(".","", dat$Preu, fixed = T)
dat$Preu <- as.numeric(as.character(dat$Preu))
```

3.1. Les dades contenen zeros o elements buits?

Les dades sí que contenen elements buits atès que els anuncis no sempre mostren la informació completa dels habitatges. En alguns casos manca saber el número d'habitacions o de lavabos.

Com gestionaries aquests casos?

La tècnica aplicada pel tractament de les dades perdudes ha estat ignorar la tupla, és a dir, no fer ús de la fila que li falten dades desaprofitant alhora la resta de dades que pugui contenir.

S'eliminen totes les files que no continguin valors

```
dat <- na.omit(dat)
```

Aquesta tècnica provocar un biaix a les dades, però el motiu d'haver ignorat les files amb camps buits és perquè si apliquéssim la tècnica de la mitjana, els valors extrems provocaria una desviació de l'estudi i un biaix encara més gran sobre l'anàlisi de les dades.

3.2. Identificació i tractament de valors extrems.

Per identificar els valors extrems utilitzem primer la comanda *summary*

```
summary(dat)
```

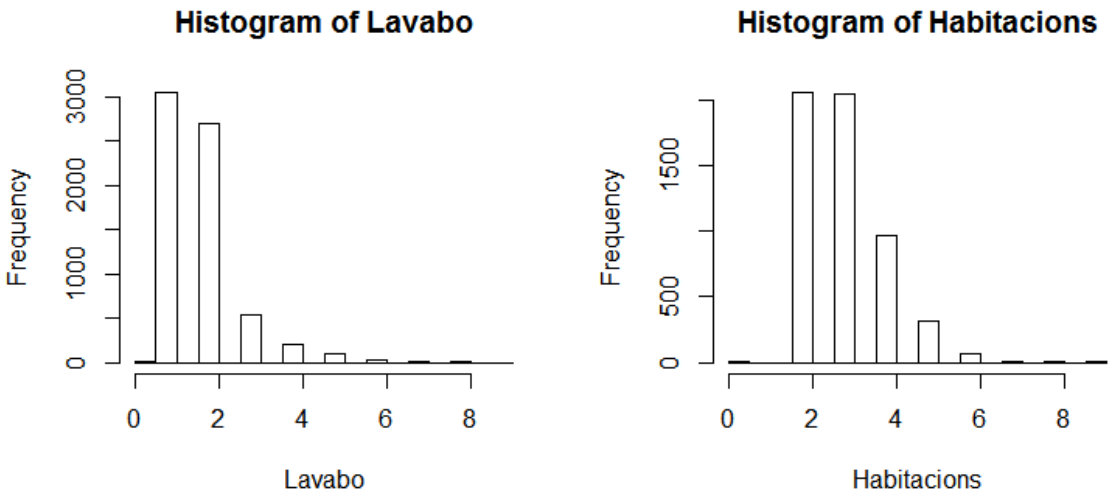
```
  Habitacions        Lavabo          Metres           Preu          TipusImmoble
Min.   :0.000   Min.   :0.000   Min.   :  1.0   1.200 €: 364   Piso       :5124
1st Qu.:2.000   1st Qu.:1.000   1st Qu.: 65.0   1.100 €: 344   Apartamento: 690
Median :3.000   Median :2.000   Median : 85.0   1.300 €: 291   Ático      : 462
Mean   :2.971   Mean   :1.742   Mean   :102.9   1.500 €: 273   Loft       : 121
3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:116.0   1.400 €: 233   Casa       : 119
Max.   :9.000   Max.   :9.000   Max.   :850.0   1.600 €: 220   Dúplex     : 103
```

Per començar, crida l'atenció que hi hagi habitatges amb 9 lavabos o que tinguin 850 metres. Executem la comanda which per saber quants i quins són els habitatges amb més de 600 metres i més de 5 lavabos

```
> which(Metres>600)
 [1]  641  642 1310 1857 2016 2368 2747 2903 2963 3182 3559 3632 4031 4186 4709 5796 6336
> which(Lavabo>5)
 [1]  109  140  641  642 1127 1433 1597 1675 2016 2149 2223 2368 2471 2683 2747 2918 2926 3475 3559 3793 3967 4031 4186 4709
[25] 5050 5164 5640 5666 5694 5964 6071 6336 6449
```

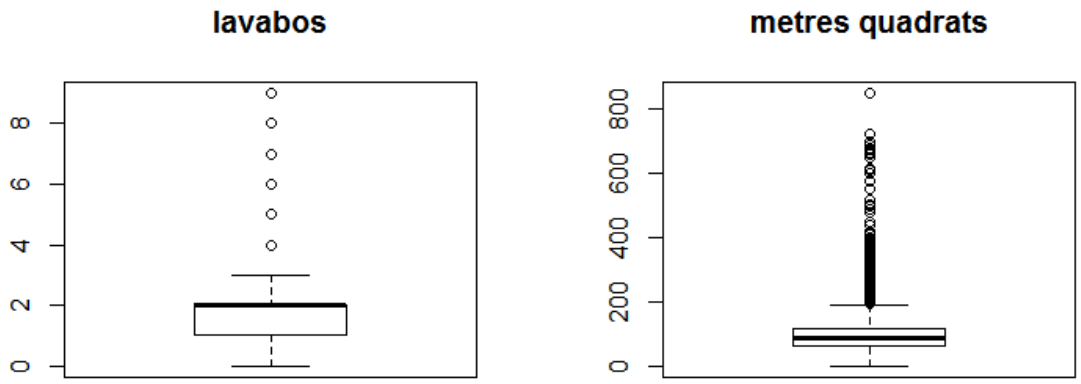
Una altre manera de detectar els valors extrems es a partir de gràfics:

```
par(mfrow=c(1,2))
hist(dat$Lavabo)
hist(dat$Habitacions)
```

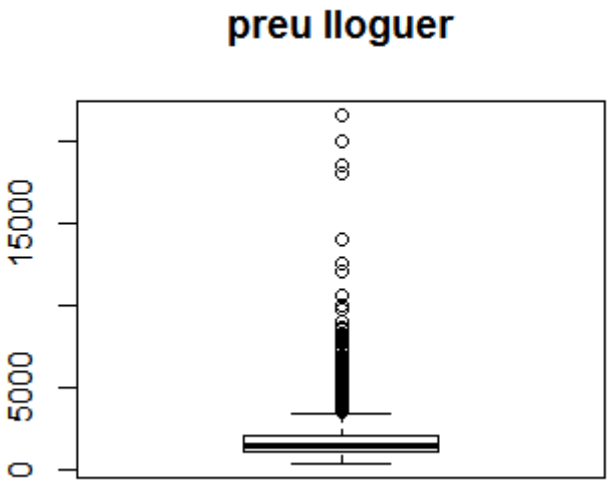


Si canviem el tipus de gràfic a diagrama de caixa podrem observar els valors inusuals en forma de punt blanc:

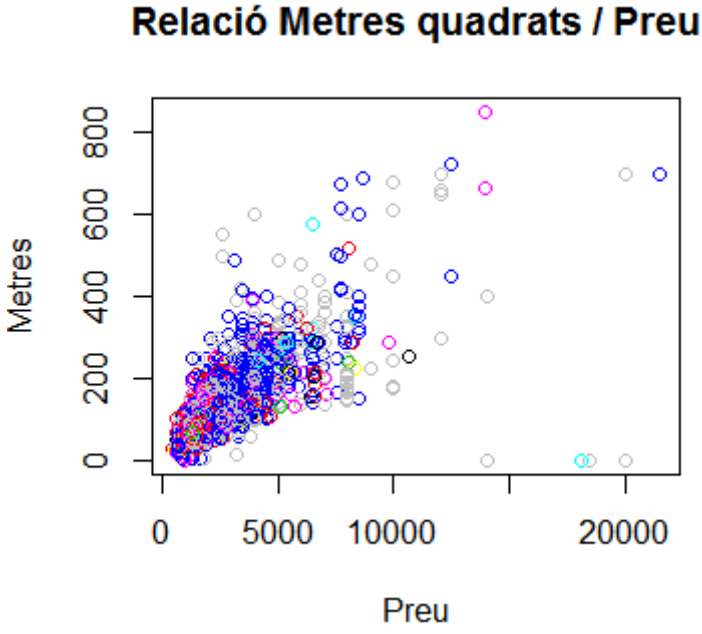
```
boxplot(dat$Lavabo,main='lavabos')
boxplot(dat$Metres,main='metres quadrats')
```




```
boxplot(as.numeric(dat$Preu),main='preu lloguer')
```




```
plot(dat$Metres~dat$Preu , col=dat$Preu,main='Relació Metres quadrats
/ Preu')
```



Podem observar que la relació metres quadrats i preu, el gruix estaria entre 0 i 400 metres i amb preu de fins a 9000 euros. D'altra banda, podem observar valors extrem a partir de 10.000 euros i habitatges amb superfícies superior a 500 metres quadrats.

Així, els valors extrems detectats no provenen d'errors de les dades, sinó possiblement perquè hi ha una gran diferencia de preu entre els habitatges del districtes, en aquest cas, el districte de Pedralbes surt fora de la mitjana però no obstant això, s'ha de tenir en compte aquestes dades i no es poden excloure.

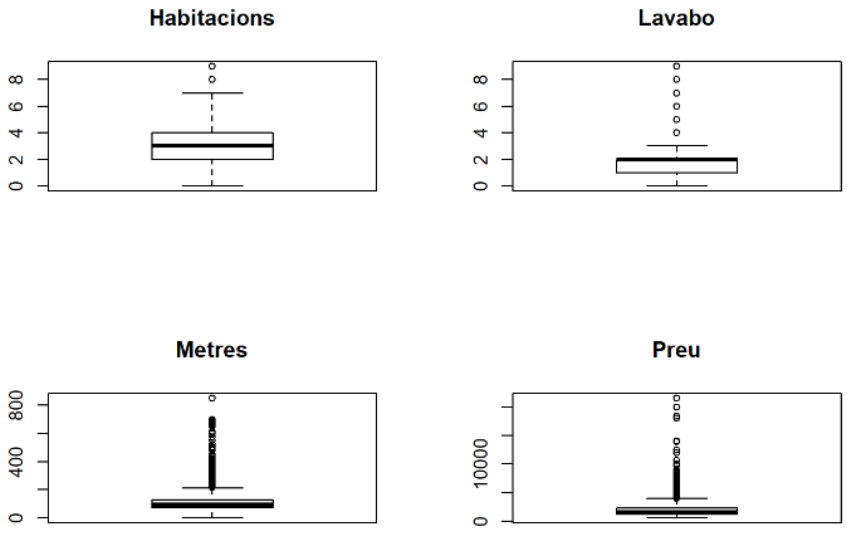
4. Anàlisi de les dades.

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

El diagrama de caixa és un mètode estandarditzat per a representar gràficament una sèrie de dades numèriques a través dels seus quartils. A continuació, s'han generat diversos diagrames de caixes per mostrar a simple vista la mitjana i els quartils de les dades, a més de representar els valors atípics.

```
install.packages("compareGroups")
require(compareGroups)
res <- compareGroups(~. - Barri, data = dat, max.xlev = 70)
restab <- createTable(res)
export2md(restab)
```

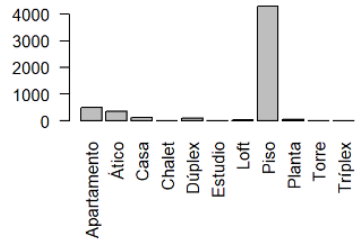
```
par(mfrow = c(2,2))
quanti <- c("Habitacions", "Lavabo", "Metres", "Preu")
for (i in seq_along(quanti)) boxplot(dat[,quanti[i]], main = quanti[i])
```



S'observa que valor atípics per totes les variables estudiades, tal i com es representa als gràfics, amb observacions per sobre del límit superior de les capses. S'obté com a observacions mitjanes aproximades, 3 habitacions, 2 lavabos, 100 metres i 1800 €.

En relació al tipus d'immoble més anunciat al web d' habitaclia per la ciutat de Barcelona, la llista l'encapçala els pisos, seguit d'apartaments i àtics.

```
plot(dat$TipusImmoble, las = 2)
```



S'ha decidit que per tal de no perdre potència a l'hora d' aplicar tests estadístics, agrupar els barris per districtes. Així doncs, passem de tenir 67 agrupacions, a tenir-ne 10, que conformaran els districtes de Barcelona:

- Sarrià-Sant Gervasi
- Les corts
- Sants-Montjuïc
- Ciutat Vella
- Eixample
- Gràcia
- Horta- Guinardó
- Nou Barris
- Sant Andreu
- Sant Martí

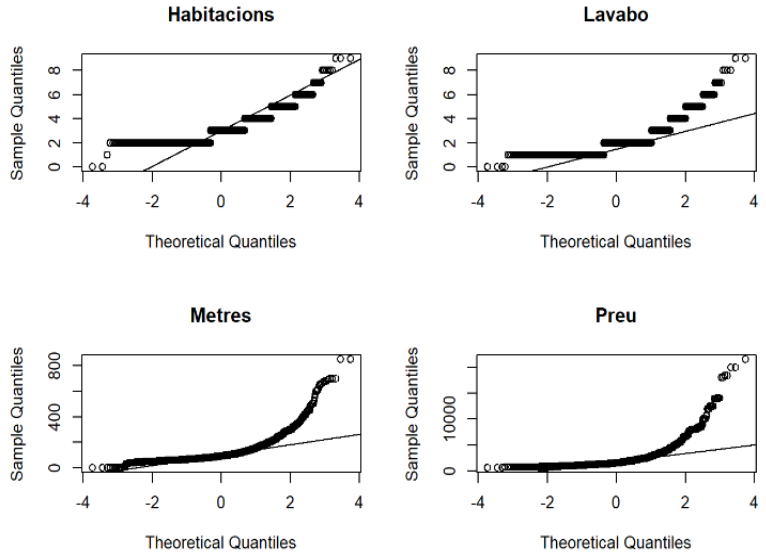
```
dat <- dat[, !names(dat) %in% "Barri"]
```

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

No s'ha pogut realitzar un test de normalitat numèric ja que el número de mostres obtingudes es lleugerament superior al límit que el test de *shapiro* permet, on es planteja com hipòtesi nul·la que una mostra prové d'una població normalment distribuïda. Aquest test és considerat un dels tests més potents per realitzar un contrast de normalitat. L'error que s'ha obtingut és: *'sample size must be between 3 and 5000'.*

```
par(mfrow = c(2,2))
for (i in seq_along(quanti)) {
  # pval <- shapiro.test(dat[,quanti[i]])$pval
  qqnorm(dat[,quanti[i]],main = quanti[i]) #main = paste0(quanti[i], "Shapiro test. P-value:", pval) )
  qqline(dat[,quanti[i]])
}
```

Per tant, passem a realitzar una avaluació gràfica de la normalitat.



Tal i com es pot observar les dades no segueixen en cap cas de forma clara la línia de distribució ajustada, per tant es podria rebutjar la hipòtesi nul·la afirmant que les dades no segueixen una distribució normal.

```
require(car)
for (i in seq_along(quanti)) print(leveneTest(dat[,quanti[i]], dat$Barri2))
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value    Pr(>F)
## group    9  42.233 < 2.2e-16 ***
##       5392
## ---
## 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  27.756 < 2.2e-16 ***
##       5392
## ---
## 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  67.371 < 2.2e-16 ***
##       5392
## ---
## 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  48.875 < 2.2e-16 ***
##       5392
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Per estudiar la homogeneïtat de les variables tenint en compte com a grups els barris, s'aplica el test de *Levene*, assumint que les variàncies de les poblacions de les quals s'extreuen diferents mostres són iguals. La prova de *Levene* avalua aquest supòsit. Es posa a prova la hipòtesi nul·la que les variàncies poblacionals són iguals.

Si el P-valor resultant de la prova de *Levene* és inferior a un cert nivell de significació (típicament 0.05), és poc probable que les diferències obtingudes en les variacions de la mostra s'hagin produït sobre la base d'un mostreig aleatori d'una població amb variàncies iguals. Per tant, la hipòtesi nul·la d'igualtat de variàncies es rebutja i es conclou que hi ha una diferència entre les variacions.

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. 5. Representació dels resultats a partir de taules i gràfiques.

- Quin tipus d'immoble és el mes publicat?

```
require(gmodels)
CrossTable(dat$TipusImmoble)
```

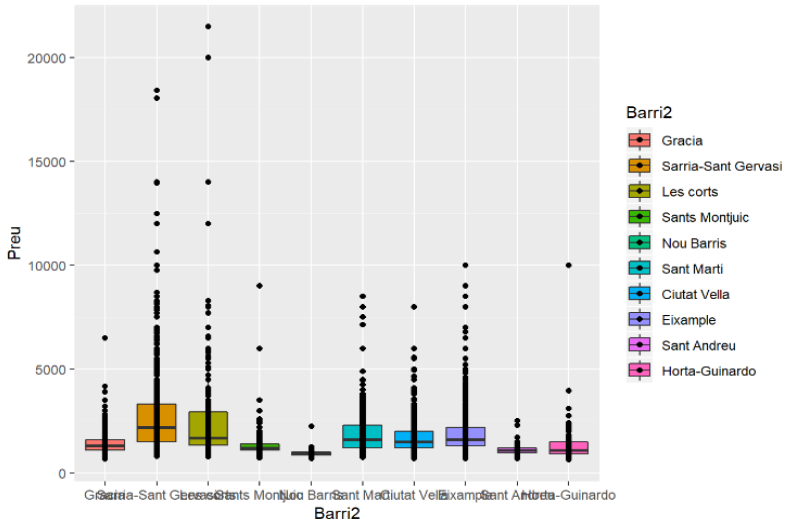
```
## Total Observations in Table:  5402
##
##
##             | Apartamento |       Ático |        Casa |      Chalet |      Dúplex |
##             |-------------|-------------|-------------|-------------|-------------|
##             |         494 |         340 |         113 |           6 |          91 |
##             |       0.091 |       0.063 |       0.021 |       0.001 |       0.017 |
##             |-------------|-------------|-------------|-------------|-------------|
##
##
##             |     Estudio |        Loft |        Piso |      Planta |       Torre |
##             |-------------|-------------|-------------|-------------|-------------|
##             |           7 |          16 |        4291 |          38 |           2 |
##             |       0.001 |       0.003 |       0.794 |       0.007 |       0.000 |
##             |-------------|-------------|-------------|-------------|-------------|
##
##
##             |     Triplex |
##             |-------------|
##             |           4 |
##             |       0.001 |
##             |-------------|
##
##
##
##
```

Tal i mostra la taula anterior, veiem que el tipus d'immoble mes publicat al portal web analitzat, és de 4291 pisos (79%), seguit de 494 apartaments. Els tipus d'immoble que menys s'ha publicat són les torres, amb un total de 2 anuncis.

Passem ara a realitzar un anàlisi descriptiu bivariant, on observarem la relació entre el Preu i els mestres quadrats del immoble.

- El preu del m2 es veu afectat per el barri on es troba l'immoble?

```
require(ggplot2)
ggplot(dat,aes(x=Barri2, y=Preu, fill=Barri2)) +
      geom_boxplot()+
      geom_point()
```



Tal i com es pot observar, del anuncis analitzats, la relació més alta a nivell de preu/m2 es troba al barri de Sarrià-Sant Gervasi, on destaquen també moltes observacions de *outliers*.

Donat que es disposa de més de dos grups, es realitza un test de *Kruskall Wallis,* que és un mètode no paramètric per a provar si un grup de dades prové de la mateixa població.

```
kruskal.test(dat$Barri2 ~dat$Preu)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  dat$Barri2 by dat$Preu
## Kruskal-Wallis chi-squared = 762.87, df = 395, p-value < 2.2e-16
```

Rebutgem doncs la hipòtesis nul·la atès que hi han diferencies de preu per barri.

A continuació, s'ajusta un model de regressió lineal per tal d'observar l'efecte que té cada barri sobre el preu:

```
(sum_mod <- summary(lm(Preu ~Barri2, data = dat)))
```

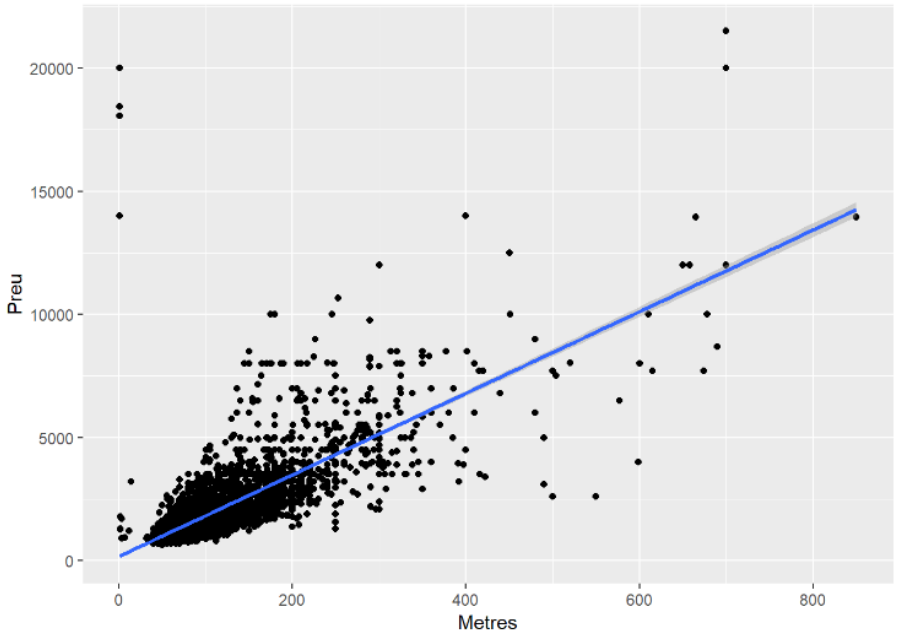
```
##
## Call:
## lm(formula = Preu ~ Barri2, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1962.0  -703.2  -281.7   255.4 18924.4
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                1444.55      72.98  19.794  < 2e-16 ***
## Barri2Sarria-Sant Gervasi  1327.44      83.96  15.810  < 2e-16 ***
## Barri2Les corts            1131.04     101.93  11.097  < 2e-16 ***
## Barri2Sants Montjuic        -82.48     117.98  -0.699  0.48451
## Barri2Nou Barris           -473.60     192.47  -2.461  0.01390 *
## Barri2Sant Marti            666.77      98.49   6.770 1.43e-11 ***
## Barri2Ciutat Vella          285.89      93.95   3.043  0.00235 **
## Barri2Eixample              458.61      81.88   5.601 2.24e-08 ***
## Barri2Sant Andreu          -323.82     155.18  -2.087  0.03696 *
## Barri2Horta-Guinardo       -141.78     124.31  -1.141  0.25409
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1469 on 5392 degrees of freedom
## Multiple R-squared:  0.1097, Adjusted R-squared:  0.1082
## F-statistic: 73.81 on 9 and 5392 DF,  p-value: < 2.2e-16
```

Tenint com a referencia el barri de Gràcia, s'observa que el barri més car (1444 € més per m2 que gràcia) és Sarrià i el més barat Nou barris (474.60€ menys per m2 que gràcia).

El percentatge de variabilitat explicada és del 10.82, és a dir, molt baix.

- El preu per m2 és veu afectat pel número de metres de l'immoble?

```
ggplot(dat, aes(x=Metres, y=Preu)) +
  geom_point()+
  geom_smooth(method=lm)
```



En aquest gràfic, s'observa que a més metres disposats a l'immoble el preu per metre quadrat augmenta. A partir de calcular el coeficient de correlació podem definir el coeficient de correlació de *Pearson* com un índex que pot utilitzar-se per a mesurar el grau de relació de dues variables sempre que ambdues siguin quantitatives i contínues.

S'haurà d'observar els següents elements:

- I) La magnitud, quant proper a 1 es troba
- II) El signe; relació directe o inversa
- III) El p.valor (tot i que es poc rellevant en correlacions)

```
cor.test(dat$Preu, dat$Metres)
```

```
##
##  Pearson's product-moment correlation
##
## data:  dat$Preu and dat$Metres
## t = 80.486, df = 5400, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7261311 0.7503880
## sample estimates:
##       cor
## 0.7384984
```

- Quin és l'augment de preu per m2?

```
(sum_mod <- summary(lm(Preu ~Metres, data = dat)))
```

```
##
## Call:
## lm(formula = Preu ~ Metres, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6680.8  -409.6  -156.7   188.9 19813.1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 170.3014    27.3161   6.234 4.88e-10 ***
## Metres       16.5646     0.2058  80.486  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1049 on 5400 degrees of freedom
## Multiple R-squared:  0.5454, Adjusted R-squared:  0.5453
## F-statistic:  6478 on 1 and 5400 DF,  p-value: < 2.2e-16
```

Es calcula el model de regressió lineal per poder observar que a la mitjana l'augment d'un metre quadrat en l'immoble fa augmentar el preu en 16€.

Donada la informació extreta d' habitaclia és vol ser capaç de predir nous pisos, per exemple el nostre. És a dir, es calcula un model multivariat que ens permeti explicar amb la màxima precisió el preu dels pisos.

```
mod_multi <- lm(Preu ~. , data =dat)
require(MASS)
stepAIC(mod_multi, trace = FALSE)
```

```
##
## Call:
## lm(formula = Preu ~ X + Habitacions + Lavabo + Metres + TipusImmoble +
##     Barri2, data = dat)
##
## Coefficients:
##               (Intercept)                          X
##                   9.41666                    0.02373
##                Habitacions                    Lavabo
##                  -68.85549                  397.46359
##                    Metres          TipusImmobleÁtico
##                  12.01147                  183.03967
##           TipusImmobleCasa         TipusImmobleChalet
##                 785.51051                  -43.77274
##         TipusImmobleDúplex        TipusImmobleEstudio
##                 -15.64471                 -349.60389
##           TipusImmobleLoft           TipusImmoblePiso
##                -401.98812                 -190.61701
##         TipusImmoblePlanta          TipusImmobleTorre
##                -197.12657                  958.64328
##        TipusImmobleTríplex  Barri2Sarria-Sant Gervasi
##                -242.51023                  223.25621
##            Barri2Les corts       Barri2Sants Montjuic
##                 273.45257                  -62.10063
##           Barri2Nou Barris           Barri2Sant Marti
##                -180.03342                  475.51974
##         Barri2Ciutat Vella              Barri2Eixample
##                 207.54188                  207.13609
##          Barri2Sant Andreu       Barri2Horta-Guinardo
##                -149.36822                 -320.64271
```

Tot i que la selecció del model a partir del criteri *d'akaike* (AIC) inclou el tipus d'immoble, un cop eliminat s'observa que es perd molt poca variabilitat explicada, per tant, decidim prescindir d'aquesta variable.

```
mod_fin <- lm(formula = Preu ~ Habitacions + Lavabo + Metres  +
    Barri2, data = dat)

##
summary(mod_fin)
```

```
##
## Call:
## lm(formula = Preu ~ Habitacions + Lavabo + Metres + Barri2, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5543.9  -419.1   -92.7   226.2 18655.1
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -101.9504    64.8625  -1.572 0.116057
## Habitacions                -81.4556    18.4572  -4.413 1.04e-05 ***
## Lavabo                     400.6301    22.5056  17.801  < 2e-16 ***
## Metres                      13.0511     0.3175  41.112  < 2e-16 ***
## Barri2Sarria-Sant Gervasi  191.6544    59.3205   3.231 0.001242 **
## Barri2Les corts            237.3360    71.0111   3.342 0.000837 ***
## Barri2Sants Montjuic       -44.3680    80.8021  -0.549 0.582963
## Barri2Nou Barris          -210.0204   131.8945  -1.592 0.111368
## Barri2Sant Marti           445.0303    67.5085   6.592 4.75e-11 ***
## Barri2Ciutat Vella         185.4782    64.7121   2.866 0.004170 **
## Barri2Eixample             171.3007    56.2446   3.046 0.002333 **
## Barri2Sant Andreu         -128.7101   106.3779  -1.210 0.226358
## Barri2Horta-Guinardo      -285.3968    85.1940  -3.350 0.000814 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1006 on 5389 degrees of freedom
## Multiple R-squared:  0.5827, Adjusted R-squared:  0.5818
## F-statistic: 627.1 on 12 and 5389 DF,  p-value: < 2.2e-16
```

A partir del model anterior fem una predicció dels nostres pisos

```
res_pred <- predict(mod_fin, newdata = data.frame(Habitacions = c(3,2),
                                    Lavabo = c(2,1),
                                    Metres = c(99,70),
                                    Barri2 = c("Nou Barris", "Gracia")))
res_pred
```

```
##        1        2
## 1536.984 1049.348
```

Per calcular el valor total del pis s'ha calculat aquest preu pels metres quadrats.

```
res_pred * c(99,70)
```

```
##         1         2
## 152161.45  73454.33
```

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

Analitzant les dades extretes del portal web habitaclia.com referents als lloguers de Barcelona publicats durant el mes d'Abril del 2019, s'han pogut respondre les preguntes plantejades a l'exercici 1. Per tant, podem concloure que:

El tipus d'immoble que més s'ha publicat correspon als anuncis de **pisos**. Aquests corresponen a un 80% del total d'anuncis de les nostres dades, mentre que la quantitat d'immobles menys publicat és de **torres**, amb un total de 2 anuncis.

El preu del immoble es veu clarament afectat per el barri on es troba. Sarrià -Sant-Gervasi, copsa el preu per metre quadrat més alt, mentre que Nou Barris té el preu més baix per metre quadrat.

El preu del immoble esta directament relacionat amb els metres quadrats dels quals disposa. Peel model lineal presentat en aquest apartat, s'ha observat un *R-square 0.54*, prou alt (dades reals) per poder afirmar aquesta relació. A més la variable Metres es mostra com estadísticament significativa.

El valor estimat de la pendent, indica que l'augment d'un metre quadrat a l'immoble, comporta una augment del preu del mateix en 16€.

7. Codi: Cal adjuntar el codi, preferiblement en R, amb el que s'ha realitzat la neteja, anàlisi i representació de les dades.

```
---
title: 'Practica 2'
author: "earinos i jlchan"
date: "`r format(Sys.time(), '%d %B, %Y')`"
output:
  html_document:
    df_print: paged
    toc: yes
    toc_float: yes
  pdf_document:
    toc: yes
editor_options:
  chunk_output_type: console
fig_caption: yes
fontsize: 10pt
geometry: margin=.95in
lang: es
link-citations: yes
linkcolor: red
number_sections: yes
csl: springer-basic-brackets.csl
always_allow_html: yes
toc: yes
urlcolor: blue
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE, message = FALSE, warning = FALSE)
```


Lectura dades

```{r}
```
```

```
 dat <- read.csv("Libro_mod.csv", sep = ";", dec = ",", fileEncoding = "Windows-1254")

library(readxl)
barris <- as.data.frame(read_excel("codis_barri.xlsx", sheet = 1))
barris_codi <- as.data.frame(read_excel("codis_barri.xlsx", sheet = 2))

# write.csv(unique(dat$Barri), file = "Barris.csv", fileEncoding = "Windows-1254", row.names = F)
```

Neteja dades

```{r}
dat$Preu <- gsub(" â, ¬","", dat$Preu)
dat$Preu <- gsub(".","", dat$Preu, fixed = T)
dat$Preu <- as.numeric(as.character(dat$Preu))
```

Podem eliminar tots els registres amb dades faltants, tot i que no es necessari, ja que si falta un valor no s'utilitzara quan es treballi amb aquella variable

```{r}
dat <- na.omit(dat)
```

Es consideren erronis els valors superior a 50.000 â, ¬ per metre quadrat

```{r}
dat$Preu[dat$Preu > 50000] <- NA
```

Identificació de valors extrems

```{r}
summary(dat)
par(mfrow=c(1,2))
hist(dat$Lavabo)
hist(dat$Habitacions)
boxplot(dat$Lavabo,main='lavabos')
```

```
boxplot(dat$Metres,main='metres quadrats')
```

Relació metres quadrats per preu

```{r}
plot(as.numeric(dat$Metres)~as.numeric(dat$Preu)  ,  col=dat$Preu,main='Relació Metres quadrats / Preu')
```

Reagrupem els barris per reduir el numero de categories

```{r}
for (i in 1:nrow(barris)) dat$Barri2[as.character(dat$Barri) == barris[i, 1]] <- barris[i, 2]

dat$Barri2 <- factor(dat$Barri2, levels = 1:10, barris_codi$Barri )
```

Descripcio dades

```{r, results='asis', eval = FALSE}
install.packages("compareGroups")
require(compareGroups)
res <- compareGroups(~. - Barri, data = dat, max.xlev = 70)
restab <- createTable(res)
export2md(restab)
```

```{r}
par(mfrow = c(2,2))
quanti <- c("Habitacions", "Lavabo", "Metres", "Preu")
for (i in seq_along(quanti)) boxplot(dat[,quanti[i]], main = quanti[i])

plot(dat$TipusImmoble, las = 2)
```

Analisis de les dades

Seleccio grups de dades...

No utlitzarem barri si no la reagrupaciÃ³ realitzada a la neteja de les dades

```{r}
dat <- dat[, !names(dat) %in% "Barri"]
```

ComprovaciÃ³ de la homogeneitat i normalitat

No es pot fer test de normalitat, n tant gran que se li suposa:
sample size must be between 3 and 5000

```{r}
par(mfrow = c(2,2))
for (i in seq_along(quanti)) {
 # pval <- shapiro.test(dat[,quanti[i]])$pval
 qqnorm(dat[,quanti[i]],main = quanti[i]) #main = paste0(quanti[i], "Shapiro test. P-value:", pval) )
 qqline(dat[,quanti[i]])
}
```

```

*Homogeneitat.*

*Es rebutja homogeneitat de les variancies per a totes les variables quantitatives, tenint en compte com a grups els barris*

```{r}
require(car)
for (i in seq_along(quanti)) print(leveneTest(dat[,quanti[i]], dat$Barri2))
```

Objectius i resoluciÃ³
```

* Quin tipus d'immoble Ã©s el mes publicat?

```{r}
require(gmodels)
CrossTable(dat$TipusImmoble)
```

<!-- * Quina relaciÃ³ de preus per m^2 quadrat hi ha a cada barri? -->

* El preu per metre quadrat es veu afectat pel barri on es troba l'immoble?

AnÃ¡lisis descriptiu bivariat

```{r}
require(ggplot2)
ggplot(dat,aes(x=Barri2, y=Preu, fill=Barri2)) +
    geom_boxplot()+
    geom_point()
```

Contrast d'hipotesis, al tenir mes de 2 grups, realizem test kruskall Wallis

```{r}
kruskal.test(dat$Barri2 ~dat$Preu)
```

Rebutjem hipotesis nula, hi han diferencies de preu per barri.

Ajustem un model de regressiÃ³ lineal per veure l'efecte de cada barri

```{r}
(sum_mod <- summary(lm(Preu ~Barri2, data = dat)))
```

Tenint com a referencia gracia, veiem que el barri mÃ©s car (1444 â, ¬ mÃ©s per m2 que grÃ cia) es Sarria i el mÃ©s barat Nou barris (474.60â, ¬ menys per m2 que grÃ cia).

El percentatge de variabilitat explicada es del `r round(sum_mod$adj.r.squared*100,2)`, Ã©s a dir, molt baix.

** El preu per m2 Ã©s veu afectat pel nÃºmero de metres de l'immoble?*

```{r}
ggplot(dat, aes(x=Metres, y=Preu)) +
  geom_point()+
  geom_smooth(method=lm)
```

A mÃ©s metres sembla q el preu per metre quadrat augmenta. Calculem el coeficient de correlaciÃ³

Ens em de fixar en:
i) la magnitud, quan proper a 1 estÃ
ii) el signe, relacio directe o inversa
iii) el p.valor (tot i que es poc rellevant en correlacions)

```{r}
cor.test(dat$Preu, dat$Metres)
```

I per Ãºltim per calcular quin es l'augment de preu per m2, per cada metre que augmenta l'immoble calculem model de regresio lineal

```{r}
(sum_mod <- summary(lm(Preu ~Metres, data = dat)))
```

** Donada la informaciÃ³ extreta d'habitaclia volem intentar ser capaÃ§os de predir nous pisos, per exemple el nostre. Ã‰s a dir, calculem un model multivariat que ens permeti explicar amb la mÃ xima precisiÃ³ el preu dels pisos.*

```{r}
mod_multi <- lm(Preu ~. , data =dat)
require(MASS)
stepAIC(mod_multi, trace = FALSE)
```

Tot i que la seleccio del model a partir del criteri d'akaike (AIC) ens inclou el tipus d'immoble, un cop eliminat veiem que perdem molt poca variabilitat explicada, per tant, decidim prescindir d'aquesta variable.

```{r}
mod_fin <- lm(formula = Preu ~ Habitacions + Lavabo + Metres  +
   Barri2, data = dat)

##
summary(mod_fin)
```

** A partir del model anterior fem una prediccIÃ³ de les nostres cases*

```{r}
res_pred <- predict(mod_fin, newdata = data.frame(Habitacions = c(3,2),
                                Lavabo = c(2,1),
                                Metres = c(99,70),
                                Barri2 = c("Nou Barris", "Gracia")))
res_pred
```

Per calcular el valor total del pis hem de calcular aquest preu pels metres quadrats.

```{r}
res_pred * c(99,70)
```

Taula de contribucions del treball.

Contribucions	Signa
Investigació prèvia	earinos, jlchan
Redacció de les respostes	earinos, jlchan
Desenvolupament codi	earinos, jlchan