Tipologia i cicle de vida de les dades
Pràctica 2 – Juny de 2020

Alumnes:

- Marc Serra Suñol
- Javier Beltran Lou

Respostes a les preguntes

Pregunta 1

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

Continuem en aquesta pràctica amb el dataset que vam obtenir a la primera pràctica de l'assignatura i que tracta sobre ofertes d'habitatges en règim de lloguer. El dataset pot ser interessant per realitzar múltiples anàlisis relacionades amb el mercat de lloguer, però en aquesta pràctica ens hem centrat en tractar de respondre una pregunta en particular: Podem fer prediccions sobre el preu que una nova oferta tindrà a partir de les dades de les ofertes del dataset?

Pregunta 2

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

En aquesta pràctica partim del dataset obtingut a la primera pràctica. En realitat, vam obtenir un nou dataset unes setmanes després d'obtenir l'original. El propòsit és el de disposar d'algunes dades actualitzades que podrien ser útils en algunes anàlisis, tot i que no és estrictament necessari per la pregunta que ens hem plantejat.

Bolquem a continuació els fragments de codi R més importants (el detall es pot veure al codi font)

```
# Càrrega del primer dataset
filePath <-"dataset_10_04_2020.csv"
ds_1004 <- read.csv(filePath, header=TRUE, sep=";", stringsAsFactors=F,
colClasses=c("Sup_m2"="character"), encoding = "UTF-8")
attach(ds_1004)
```

```
# Càrrega del segon dataset
filePath2 <-"dataset_26_04_2020.csv"
ds_2604 <- read.csv(filePath2, header=TRUE, sep=";", stringsAsFactors=F,
colClasses=c("Sup_m2"="character"), encoding = "UTF-8")
attach(ds_2604)
```

(…)

Dins del concepte de selecció de dades, i també com una operació de neteja més, es va considerar haver d'esborrar duplicitats.

```
# Elements duplicats al primer dataset
dup_1004 <- ds_1004$Id[duplicated(ds_1004$Id)]
```

```
# Elements duplicats al segon dataset
dup_2604 <- ds_2604$Id[duplicated(ds_2604$Id)]


# Seleccionant els duplicats i ordenant per id podem explorar l'aspecte de
les files:
# Primer dataset
duplicated_rows_1004 <- filter(ds_1004, Id %in% c(dup_1004))
duplicated_rows_1004 <- duplicated_rows_1004[with(duplicated_rows_1004,
order(duplicated_rows_1004$Id)),]
#duplicated_rows_1004

# Segon dataset
duplicated_rows_2604 <- filter(ds_2604, Id %in% c(dup_2604))
duplicated_rows_2604 <- duplicated_rows_2604[with(duplicated_rows_2604,
order(duplicated_rows_2604$Id)),]
#duplicated_rows_2604


# Després d'una exploració (accedint a la font original) d'alguns dels valors
duplicats, s'arriba a la conclusió que es tracta d'un problema en el scraper.
Es pot eliminar amb seguretat els registres duplicats i és així com es
procedeix:

# Eliminació de duplicitats:
ds_1004 <- ds_1004[!duplicated(ds_1004$Id), ]
ds_2604 <- ds_2604[!duplicated(ds_2604$Id), ]
```

(...)

Nota: Al codi font es comprova també que, efectivament, els duplicats han desaparegut

Finalment es realitza una integració dels dos datasets:

```
# En realitat, del segon dataset només ens interessa conservar el preu i el
id per fer un criteri de merge, doncs la resta de camps (excepte la data) són
iguals
# PreuActual
ds_2604 <- select(ds_2604, "Id", "PreuActual")

# Renombrem la columna preu al primer dataset
ds_1004 <- rename(ds_1004, Preu=PreuActual)

# Ara sí, el merge:
dataset = merge(x = ds_1004, y = ds_2604, by = "Id", all = TRUE)

# No ens interessa les ofertes que eren noves al segon dataset
dataset <- dataset[!is.na(dataset$idx),]
```

I en aquest punt, “dataset” ja és el dataset amb què treballarem a la resta del codi. Lògicament, cal fer encara molta feina de neteja

Pregunta 3

Neteja de les dades

```
# Llistar les columnes
colnames(dataset)
```

```
 [1] "Id"                   "idx"                  "Url"
 [4] "TipusOferta"          "TipusImmoble"         "Municipi"
 [7] "Provincia"            "Zona"                 "Preu"
[10] "DataOferta"           "Sup_m2"               "NumHabitacions"
[13] "NumBanys"             "AnyConstruccio"       "Planta"
[16] "Parking"              "Calefaccio"           "AC"
[19] "Moblat"               "EficienciaEnergetica" "ClasseEmissions"
[22] "Jardi"                "Ascensor"             "PreuActual"
```

Durant el procés de neteja, alguns dels atributs s'eliminaran. Inicialment, prescindim ja d'alguns d'ells:

```
dataset <- select(dataset, -c(Id, Provincia, Municipi, idx, DataOferta,
TipusOferta))
```

(...)

Una mica d'acondicionament per treballar amb els tipus més adients

```
# Variables categòriques
dataset$TipusImmoble <- as.factor(dataset$TipusImmoble)
dataset$Zona <- as.factor(dataset$Zona)
dataset$EficienciaEnergetica <- as.factor(dataset$EficienciaEnergetica)
dataset$ClasseEmissions <- as.factor(dataset$ClasseEmissions)

# Variables booleanes
dataset$Parking <- as.logical(dataset$Parking)
dataset$Calefaccio <- as.logical(dataset$Calefaccio)
dataset$AC <- as.logical(dataset$AC)
dataset$Moblat <- as.logical(dataset$Moblat)
dataset$Jardi <- as.logical(dataset$Jardi)
dataset$Ascensor <- as.logical(dataset$Ascensor)

# Posar les superfícies com a enters:
dataset$Sup_m2 = sub("\\.", "", dataset$Sup_m2)
dataset$Sup_m2 = as.integer(dataset$Sup_m2)
```

(...)

Es va descobrir alguna informació una mica incoherent

```
unique(dataset$EficienciaEnergetica)
unique(dataset$ClasseEmissions)
```

```
[1] G E F D B   A C Z
Levels:  A B C D E F G Z
 [1] G E F D B   A C f e g d
Levels:  A B C d D e E f F g G
```

Duplicitats que es van resoldre així

```
dataset$ClasseEmissions <-  as.factor(toupper(dataset$ClasseEmissions))
unique(dataset$ClasseEmissions)
```

Tot i que finalment aquests atributs no van ser rellevants en les nostres anàlisis..

(...)

Sobre els elements buits

Trobem que hi ha diferents atributs que presenten “missing values”:

```
summary(dataset)
```

```
Url            TipusImmoble                      Zona       PreuInicial
 Length:8008        atico : 387                         : 652   Min.   :  350
 Class :character   casa  : 852   Dreta de l´Eixample   : 536   1st Qu.:  990
 Mode  :character   duplex: 235   Centre                : 506   Median : 1300
                    otros : 400   Sant Gervasi - Galvany: 407   Mean   : 1801
                    piso  :6134   Gòtic                 : 276   3rd Qu.: 1950
                                  Raval                 : 267   Max.   :35500
                                  (Other)               :5364   NA's   :10
     Sup_m2       NumHabitacions       NumBanys      AnyConstruccio
 Min.   :   2.0   Min.   :  1.000   Min.   : 2.000   Min.   :   1
 1st Qu.:  67.0   1st Qu.:  2.000   1st Qu.: 2.000   1st Qu.:1960
 Median :  90.0   Median :  3.000   Median : 2.000   Median :1977
 Mean   : 123.1   Mean   :  2.971   Mean   : 2.534   Mean   :1968
 3rd Qu.: 130.0   3rd Qu.:  4.000   3rd Qu.: 3.000   3rd Qu.:2003
 Max.   :2600.0   Max.   :850.000   Max.   :18.000   Max.   :2020
 NA's   :28       NA's   :128       NA's   :3630     NA's   :4050
     Planta         Parking        Calefaccio          AC
 Min.   : 1.000   Mode :logical   Mode :logical   Mode :logical
 1st Qu.: 1.000   FALSE:5616      FALSE:2030      FALSE:3149
 Median : 3.000   TRUE :2392      TRUE :5978      TRUE :4859
 Mean   : 3.073
 3rd Qu.: 4.000
 Max.   :25.000
 NA's   :3454
   Moblat        EficienciaEnergetica ClasseEmissions   Jardi
 Mode :logical   G      :4194         G      :3961    Mode :logical
 FALSE:3951      E      :1735         E      :1767    FALSE:8008
 TRUE :4057              : 687                :  687
                 D      : 482         D      : 597
                 F      : 320         F      : 386
                 C      : 252         C      : 215
                 (Other): 338         (Other): 395
  Ascensor         PreuActual    HaEstatLlogat     PreuMetre2
 Mode :logical   Min.   :  350   Mode :logical   Min.   :  0.913
 FALSE:2694      1st Qu.:  980   FALSE:6855      1st Qu.: 11.723
 TRUE :5314      Median : 1300   TRUE :1153      Median : 15.048
                 Mean   : 1795                   Mean   : 16.409
                 3rd Qu.: 1950                   3rd Qu.: 19.286
                 Max.   :35500                   Max.   :700.000
                 NA's   :9                       NA's   :36
```

Es va decidir provar diferents tipus de tractaments per valors buits que hem vist en l’assignatura. A continuació, alguns exemples:

Registres que no compten amb alguns del preus (inicial i actual): Vam optar per eliminar registres, doncs van resultar ser pocs i no ens ajuden a respondre la pregunta plantejada

```
dataset <- dataset[!is.na(dataset$PreuInicial),]
dataset <- dataset[!is.na(dataset$PreuActual),]
```

En alguns altres casos optem per l'aproximació basada en imputació amb mesures de valors centrals. Per exemple, per als missing vàlues de "NumHabitacions" optem per imputar la mediana

```
dataset$NumHabitacions[is.na(dataset$NumHabitacions)] <-
median(dataset$NumHabitacions, na.rm = TRUE)
```

I per altres valors NAs es va optar per la técnica més avançada d'imputació mitjançant els k nearest neighbors.

Per exemple, per "AnyConstrcuccio" es segueix la següent lògica:

```
library(VIM)

# El drama de kNN és escollir el valor de k. Per fer-ho bé, podríem entrenar
un subconjunt del dataset (amb valors no NA) i veure amb quina k podem predir
# millor la resta de registres que no han participat al training. En aquest
punt ens coformarem amb uns quants valor de k i veure en quin cas les mesures
# centrals no s'allunyen massa de les originals
ds_knn_k5 <- kNN(dataset, variable = c("AnyConstruccio"), dist_var =
c("TipusImmoble", "Zona", "PreuInicial", "Sup_m2", "NumHabitacions",
"NumBanys", "Planta", "Parking", "Calefaccio", "AC", "Moblat",
"EficienciaEnergetica", "ClasseEmissions", "Jardi", "Ascensor"))
ds_knn_k6 <- kNN(dataset, variable = c("AnyConstruccio"), dist_var =
c("TipusImmoble", "Zona", "PreuInicial", "Sup_m2", "NumHabitacions",
"NumBanys", "Planta", "Parking", "Calefaccio", "AC", "Moblat",
"EficienciaEnergetica", "ClasseEmissions", "Jardi", "Ascensor"), k = 6)
ds_knn_k15 <- kNN(dataset, variable = c("AnyConstruccio"), dist_var =
c("TipusImmoble", "Zona", "PreuInicial", "Sup_m2", "NumHabitacions",
"NumBanys", "Planta", "Parking", "Calefaccio", "AC", "Moblat",
"EficienciaEnergetica", "ClasseEmissions", "Jardi", "Ascensor"), k = 15)
ds_knn_k50 <- kNN(dataset, variable = c("AnyConstruccio"), dist_var =
c("TipusImmoble", "Zona", "PreuInicial", "Sup_m2", "NumHabitacions",
"NumBanys", "Planta", "Parking", "Calefaccio", "AC", "Moblat",
"EficienciaEnergetica", "ClasseEmissions", "Jardi", "Ascensor"), k = 50)

summary(dataset$AnyConstruccio)
summary(ds_knn_k5$AnyConstruccio)
summary(ds_knn_k6$AnyConstruccio)
summary(ds_knn_k15$AnyConstruccio)
summary(ds_knn_k50$AnyConstruccio)
```

Finalment:

```
dataset <- kNN(dataset, variable = c("AnyConstruccio"), dist_var =
c("TipusImmoble", "Zona", "PreuInicial", "Sup_m2", "NumHabitacions",
"NumBanys", "Planta", "Parking", "Calefaccio", "AC", "Moblat",
"EficienciaEnergetica", "ClasseEmissions", "Jardi", "Ascensor"))
```

I per l'atribut "Planta" s'actua del mateix mode:

```
# Imputar amb kNN:
ds_knn_k5 <- kNN(dataset, variable = c("Planta"), dist_var =
c("TipusImmoble", "Zona", "PreuInicial", "Sup_m2", "NumHabitacions",
"NumBanys", "AnyConstruccio", "Parking", "Calefaccio", "AC", "Moblat",
"EficienciaEnergetica", "ClasseEmissions", "Jardi", "Ascensor"))
```

```r
ds_knn_k6 <- kNN(dataset, variable = c("Planta"), dist_var =
c("TipusImmoble", "Zona", "PreuInicial", "Sup_m2", "NumHabitacions",
"NumBanys", "AnyConstruccio", "Parking", "Calefaccio", "AC", "Moblat",
"EficienciaEnergetica", "ClasseEmissions", "Jardi", "Ascensor"), k = 6)
ds_knn_k15 <- kNN(dataset, variable = c("Planta"), dist_var =
c("TipusImmoble", "Zona", "PreuInicial", "Sup_m2", "NumHabitacions",
"NumBanys", "AnyConstruccio", "Parking", "Calefaccio", "AC", "Moblat",
"EficienciaEnergetica", "ClasseEmissions", "Jardi", "Ascensor"), k = 15)
ds_knn_k50 <- kNN(dataset, variable = c("Planta"), dist_var =
c("TipusImmoble", "Zona", "PreuInicial", "Sup_m2", "NumHabitacions",
"NumBanys", "AnyConstruccio", "Parking", "Calefaccio", "AC", "Moblat",
"EficienciaEnergetica", "ClasseEmissions", "Jardi", "Ascensor"), k = 50)

summary(dataset$Planta)
summary(ds_knn_k5$Planta)
summary(ds_knn_k6$Planta)
summary(ds_knn_k15$Planta)
summary(ds_knn_k50$Planta)

# Ens quedem amb k = 5
dataset <- kNN(dataset, variable = c("Planta"), dist_var = c("TipusImmoble",
"Zona", "PreuInicial", "Sup_m2", "NumHabitacions", "NumBanys",
"AnyConstruccio", "Parking", "Calefaccio", "AC", "Moblat",
"EficienciaEnergetica", "ClasseEmissions", "Jardi", "Ascensor"))
```

Cas especial és el tractament que fem amb les zones, doncs formaran part de la nostra anàlisi posterior:

````
```{r Neteja_Zones}

# Primer de tot mirarem si hi han camps buits
sum(is.na(dataset$Zona))
sum(dataset$Zona == "")

# Veiem que hi ha 639 casos amb la variable buida, aquests no els tindrem en
compte
dataset <- filter(dataset, Zona != "")

# Mirem quantes zones diferents hi ha
length(unique(dataset$Zona))

# Mirem la representativitat de cada zona
zones <- table(dataset$Zona)

# Ens quedarem nomes amb els casos que tinguin un minim de representativitat,
en el nostre cas agafarem nomes les zones amb mes de 50 apartaments en
lloguer
zones_a_agafar <- names(zones[zones>=50])
dataset <- filter(dataset, Zona %in% zones_a_agafar)

```
````

Com es pot veure en el fragment de codi anterior, es va decidir treballar únicament amb les zones amb un mínim de representativitat. Concretament aquelles de què disposem d¡almenys 50 ofertes de lloguer.

Sobre els valors extrems

En el codi de la pràctica s'estudien els outliers de les diferents variables i es prenen decisions al respecte (són valors legítims?, Es tracta d'errors en les dades?, ...).

Tornant al summary, ja mostrat anteriorment:

```
summary(dataset)
```

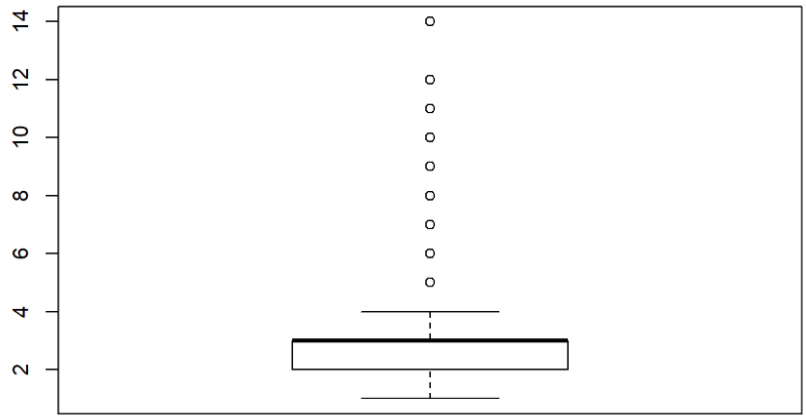
```
Url             TipusImmoble                      Zona        PreuInicial
 Length:8008        atico : 387                          : 652   Min.   :  350
 Class :character   casa  : 852   Dreta de l´Eixample    : 536   1st Qu.:  990
 Mode  :character   duplex: 235   Centre                 : 506   Median : 1300
                    otros : 400   Sant Gervasi - Galvany: 407   Mean   : 1801
                    piso  :6134   Gòtic                  : 276   3rd Qu.: 1950
                                  Raval                  : 267   Max.   :35500
                                  (Other)                :5364   NA's   :10
     Sup_m2        NumHabitacions       NumBanys      AnyConstruccio
 Min.   :   2.0   Min.   :  1.000   Min.   : 2.000   Min.   :   1
 1st Qu.:  67.0   1st Qu.:  2.000   1st Qu.: 2.000   1st Qu.:1960
 Median :  90.0   Median :  3.000   Median : 2.000   Median :1977
 Mean   : 123.1   Mean   :  2.971   Mean   : 2.534   Mean   :1968
 3rd Qu.: 130.0   3rd Qu.:  4.000   3rd Qu.: 3.000   3rd Qu.:2003
 Max.   :2600.0   Max.   :850.000   Max.   :18.000   Max.   :2020
 NA's   :28       NA's   :128       NA's   :3630     NA's   :4050
     Planta         Parking        Calefaccio          AC
 Min.   : 1.000   Mode :logical   Mode :logical   Mode :logical
 1st Qu.: 1.000   FALSE:5616      FALSE:2030      FALSE:3149
 Median : 3.000   TRUE :2392      TRUE :5978      TRUE :4859
 Mean   : 3.073
 3rd Qu.: 4.000
 Max.   :25.000
 NA's   :3454
   Moblat        EficienciaEnergetica ClasseEmissions   Jardi
 Mode :logical   G      :4194         G      :3961    Mode :logical
 FALSE:3951      E      :1735         E      :1767    FALSE:8008
 TRUE :4057              : 687                : 687
                 D      : 482         D      : 597
                 F      : 320         F      : 386
                 C      : 252         C      : 215
                 (Other): 338         (Other): 395
  Ascensor         PreuActual    HaEstatLlogat     PreuMetre2
 Mode :logical   Min.   :  350   Mode :logical   Min.   :  0.913
 FALSE:2694      1st Qu.:  980   FALSE:6855      1st Qu.: 11.723
 TRUE :5314      Median : 1300   TRUE :1153      Median : 15.048
                 Mean   : 1795                   Mean   : 16.409
                 3rd Qu.: 1950                   3rd Qu.: 19.286
                 Max.   :35500                   Max.   :700.000
                 NA's   :9                       NA's   :36
```

Veiem com amb una simple anàlisi exploratòria ja detectem alguns outliers potencials que s'han destacat en vermell. Per exemple, sembla segur que no pot haver-hi un immoble construït l'any 1. També sembla complicat que un immoble en lloguer disposi de 850 habitacions.

Utilitzant boxplots podem veure per cada variable per separat com es distribueixen els valors i si tenim outliers. Posteriorment es decideix què fer amb ells segons si es consideren legítims (es mantenen els valors) o no (s'eliminen del dataset)

Variable "NumHabitacions":

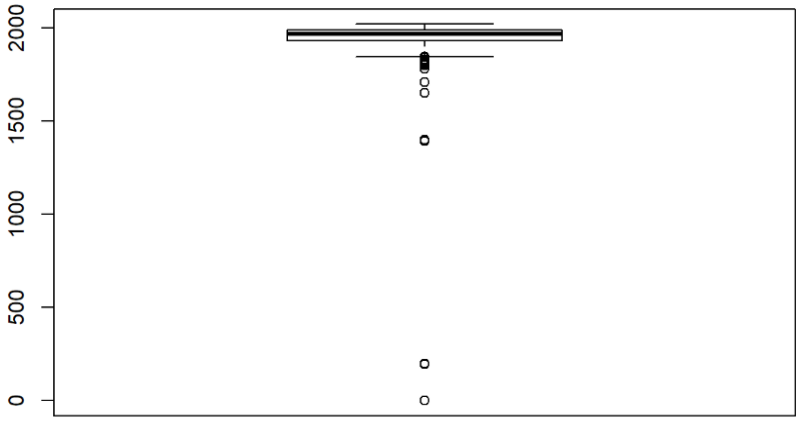
```
boxplot(dataset$NumHabitacions)
```



Sorprenen els habitatges de més de 5 habitacions (quelcom detectat també amb el criteri del boxplot). Una exploració a les fonts d'informació original va posar de manifest que es tractava, però, de dades correctes, al tractar-se d'ofertes pensades per ubicar centres educatius i similars.

Variable "AnyConstrucció"

```
boxplot(dataset$AnyConstruccio)
```

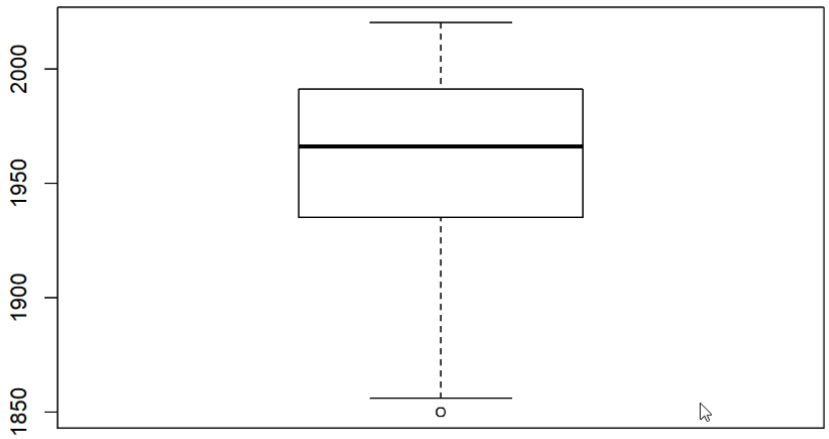


Es va trobar alguns errors evidents (anys de construcció pre-descobriment d'Amèrica) i alguns valors legítims però d'edificacions realment antigues. Es va decidir no tenir en consideració edificis massa antics:

```
# EN la gràfica anterior veiem que hi ha molts possibles outliers degut a
habitatges antics

ds_modernista <- filter(dataset, AnyConstruccio < 1900 &
!is.na(AnyConstruccio))
# Al filtrar pels més antics del 1900 ens donem compte que a Bcn hi ha
bastants habitatges del segle XIX i fins i tot del XVIII que són correctes

ds_premodernista <- filter(dataset, AnyConstruccio < 1850 &
!is.na(AnyConstruccio))

# Observem (consultant la font original) que encara hi ha alguns immobles
vàlids anteriors al 1850, de totes formes decidim esborrar-los i, de pas,
eliminar també alguns registres que amb seguretat no són correctes
```

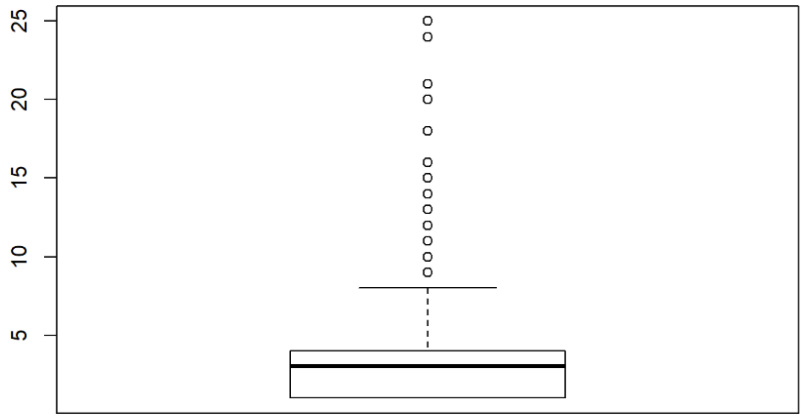
```r
dataset <- filter(dataset, AnyConstruccio >= 1850 | is.na(AnyConstruccio))

# Anem a veure si ara surt un boxplot menys infumable
boxplot(dataset$AnyConstruccio)
```

El nou boxplot té millor pinta:



Variable “Planta”:

```r
boxplot(dataset$Planta)
```



Sembla que els edificis amb ofertes de lloguer amb més de 10 plantes són bastant excepcionals a l’àrea de Barcelona.

```r
# Sembla que els pisos on la planta estan per sobre de la 8 no són molt
populars a Bcn
ds_gratacels_de_bcn_que_serien_poca_cosa_a_manhattan <-
dataset[dataset$Planta >= 10 & !is.na(dataset$Planta),]
```

Després d’investigar-los, es va trobar que són ofertes legítimes i no es va dur a terme cap operació extra.

Pregunta 4

Anàlisi de les dades.

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

Dins d'aquesta secció, i preparant ja la fase analítica, anem a crear noves variables a mode de grups, tot i que és possible que no tots els acabem fent servir. Podem fer servir:

- Zones: Ens podem fixar en algunes de les zones amb més ofertes
- Tipus d'oferta: Pot resultar interessant agrupar les ofertes segons si són cases, pisos, àtics, etc...
- Franja de m2: Podem discretitzar d'alguna manera la variable dels metres quadrats per distingir 3 categories: habitatges petits, mitjans i grans.
- Nombre d'habitacions

````
+ Zones:
```{r Grup_Zones}
# Com a exemple crearem els d'algunes de les zones amb mes apartaments en
lloguer:
dataset.Zones.SantGervasi <- filter(dataset, Zona == "Sant Gervasi -
Galvany")
dataset.Zones.Centre <- filter(dataset, Zona == "Centre")
dataset.Zones.DretaEixample <- filter(dataset, Zona == "Dreta de l´Eixample")
```
+ Tipus d'apartament:
```{r Grup_Tipus}
dataset.Tipus.Pis <- filter(dataset, TipusImmoble == "piso")
dataset.Tipus.Casa <- filter(dataset, TipusImmoble == "casa")
dataset.Tipus.Atic <- filter(dataset, TipusImmoble == "atico")
dataset.Tipus.Duplex <- filter(dataset, TipusImmoble == "duplex")
dataset.Tipus.Altres <- filter(dataset, TipusImmoble == "otros")
```
+ Franja de m2:
```{r Grup_Tamany}
dataset.Tamany.Petit <- filter(dataset, Sup_m2 < 75)
dataset.Tamany.Mitja <- filter(dataset, Sup_m2 >= 75 & Sup_m2 <= 105)
dataset.Tamany.Gran <- filter(dataset, Sup_m2 > 105)
```
+ Numero d'habitacions:
```{r Grup_Habitacions}
dataset.Habitacions.1 <- filter(dataset, NumHabitacions == 1)
dataset.Habitacions.2 <- filter(dataset, NumHabitacions == 2)
dataset.Habitacions.3 <- filter(dataset, NumHabitacions == 3)
dataset.Habitacions.4 <- filter(dataset, NumHabitacions == 4)
dataset.Habitacions.5_o_mes <- filter(dataset, NumHabitacions >= 5)
```
````

Comprovació de la normalitat i homogeneïtat de la variància.

En aquest punt, s'utilitza el test de Shapiro-Wilk per tal de comparar la distribució de les dades amb una distribució normal. Assumirem, com a hipòtesi nul·la, que la població està distribuïda normalment, si el p-valor és més petit que el nivell de significació, en el nostre cas farem servir un valor típic $\alpha=0.05$, llavors la hipòtesi nul·la és rebutjada i es conclou que les dades no compten amb una distribució normal.

```{r Shapiro_Wilk}
alpha <- 0.05

cat("La variable PreuActual no segueix una distribuio normal per les seguents zones:\n")
for (i in 1:length(zones_a_agafar)){
  zona <- zones_a_agafar[i]
  sResult <- shapiro.test(dataset[dataset$Zona  == zona,"PreuActual"])
  if(sResult$p.value < alpha){
    cat(zona)
  }
}


cat("\n\nLa variable Sup_m2 no segueix una distribuio normal per les seguents zones:\n")
for (i in 1:length(zones_a_agafar)){
  zona <- zones_a_agafar[i]
  sResult <- shapiro.test(dataset[dataset$Zona  == zona,"Sup_m2"])
  if(sResult$p.value < alpha){
    cat(zona)
  }
}


cat("\n\nLa variable NumHabitacions no segueix una distribuio normal per les seguents zones:\n")
for (i in 1:length(zones_a_agafar)){
  zona <- zones_a_agafar[i]
  sResult <- shapiro.test(dataset[dataset$Zona  == zona,"NumHabitacions"])
  if(sResult$p.value < alpha){
    cat(zona)
  }
}
```

Els resultats del test són:

```
La variable PreuActual no segueix una distribuio normal per les seguents zones:

BarcelonetaCamp d´en Grassot - Gràcia N.Camp de l´ArpaCentreDiagonal Mar - La Mar BellaDreta de l´EixampleEsquerra Alta de l´EixampleEsquerra Baixa de l´EixampleFort PiencGava MarGlòries El ParcGòticLes CortsPedralbesPoble SecPoblenouPutget - FarróRavalSagrada FamíliaSant AntoniSant Gervasi - BonanovaSant Gervasi - GalvanySantsSarriàSt. Pere - Sta. Caterina - El BornTres TorresVallcarca - PenitentsVila de Gràcia

La variable Sup_m2 no segueix una distribuio normal per les seguents zones:

BarcelonetaCamp d´en Grassot - Gràcia N.CentreDiagonal Mar - La Mar BellaDreta de l´EixampleEsquerra Alta de l´EixampleEsquerra Baixa de l´EixampleGava MarGlòries El ParcGòticLes CortsPedralbesPoble SecPoblenouPutget - FarróRavalSant AntoniSant Gervasi - BonanovaSant Gervasi - GalvanySant Ramon - MaternitatSantsSarriàSt. Pere - Sta. Caterina - El BornTres TorresVallcarca - PenitentsVila de Gràcia

La variable NumHabitacions no segueix una distribuio normal per les seguents zones:
```

```
BarcelonetaCamp d´en Grassot - Gràcia N.Camp de l´ArpaCentreDiagonal Mar - La Mar
BellaDreta de l´EixampleEsquerra Alta de l´EixampleEsquerra Baixa de l´EixampleFort
PiencGava MarGlòries El ParcGòticLes CortsPedralbesPoble SecPoblenouPutget -
FarróRavalSagrada FamíliaSant AntoniSant Gervasi - BonanovaSant Gervasi - GalvanySant
Ramon - MaternitatSantsSarriàSt. Pere - Sta. Caterina - El BornTres TorresVallcarca -
PenitentsVila de Gràcia
```

Veiem que en la gran majoria de casos no segueix una distribució normal i, per tant, per mirar la homogeneïtat de les variàncies farem servir el test de Fligner-Killeen. La hipòtesi nul·la assumeix igualtat de variàncies en els diferents grups de dades, de manera que p-valors inferiors al nivell de significació indicaran heteroscedasticitat.

```
```{r Fligner_Killeen}
fligner.test(PreuActual ~ Zona, data = dataset)
```
```

Els resultats del test són:

```
Fligner-Killeen test of homogeneity of variances

data:  PreuActual by Zona
Fligner-Killeen:med chi-squared = 998.42, df = 28, p-value < 2.2e-16
```

Com veiem, p-value < 0.05, lo qual indica heteroscedasticitat.

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.

Primer de tot farem una prova de correlació entre les variables numèriques per a veure quines variables poden estar relacionades amb el preu de l'apartament.

```
```{r Probes_Estadistiques_1}
#1 Fem correlacio amb del Preu per a la resta de variables numeriques, per
veure quines variables afecten mes en el preu final(nomes per variables
numeriques)
vars <- c("Sup_m2", "NumHabitacions", "NumBanys", "AnyConstruccio", "Planta"
)
xy <- matrix(NA, nrow = length(vars), ncol = 3)

i <- 1

for (var in vars){
  result <- cor.test(dataset[,"PreuActual"], dataset[,var])
  xy[i, ] <- c(var, result$estimate, result$p.value)
  i <- i + 1
}

colnames(xy) <- c("Var", "cor", "p.value")
xy <- as.data.frame(xy)
xy
```
```

Taula resultant:

Sup_m2	0.765699247428024

NumHabitacions	0.535923263129118
NumBanys	0.68251672397594
AnyConstruccio	0.111612787480778
Planta	0.105649975008718

A la taula resultant veiem que de les diferents variables numèriques del nostre dataset, per al cas del preu actual la variable mes significativa es la de la superfície del apartament a llogar, seguit del nombre de banys. En els dos casos p.value es prou baix per a que puguem prendre aquestes tendències com a fiables. De fet el que indica la taula es que en dos apartaments similars, si tota la resta de variables son iguals, però un es una mica mes gran que l'altre, el mes gran serà mes car.

De tota manera, podríem mirar si el nombre de habitacions o el nombre de banys tindrà una relació directa amb el tamany de l'apartament, que sembla lògic pensar que pot ser així:

```
```{r Probes_Estadistiques_1_2}
#1 Fem correlacio amb del Preu per a la resta de variables numeriques, per
veure quines variables afecten mes en el preu final(nomes per variables
numeriques)
vars <- c("NumHabitacions", "NumBanys")
xy <- matrix(NA, nrow = length(vars), ncol = 3)

i <- 1

for (var in vars){
  result <- cor.test(dataset[,"Sup_m2"], dataset[,var])
  xy[i, ] <- c(var, result$estimate, result$p.value)
  i <- i + 1
}

colnames(xy) <- c("Var", "cor", "p.value")
xy <- as.data.frame(xy)
xy
```
```

Var	**Cor**	**p.value**
NumHabitacions	0.669612539198053	0
NumBanys	0.717447845564508	0

Com es pot veure a la taula, sí que tenen una gran relació entre si.

Tot seguit farem una prova de regressió lineal per veure si seriem capaços de predir, per a un apartament amb certes característiques (tamany, zona i si esta moblat o no), el preu per al que el podríem llogar. Aquesta és una de les preguntes principals que ens havíem plantejat ja des de la primera pràctica:

```
```{r Probes_Estadistiques_2}
#2 Regressio Linial
lm <- lm(PreuActual ~ Sup_m2 + Zona + Moblat, data=dataset)
#lm$coefficients
summary(lm)
```
```

```

Resultats:

```
Residuals:
     Min   1Q   Median   3Q   Max
-18979.0   -341.4   -88.5   203.4 15940.4

Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)   582.1941   99.9026   5.828 6.00e-09 ***
Sup_m2   12.3726   0.1669 74.132 < 2e-16 ***
ZonaCamp d´en Grassot - Gràcia N.   -124.2099   148.1400 -0.838 0.401813
ZonaCamp de l´Arpa   -185.9180   149.8237 -1.241 0.214700
ZonaCentre   -610.5560   109.2034 -5.591 2.39e-08 ***
ZonaDiagonal Mar - La Mar Bella   875.1998   154.5952   5.661 1.59e-08 ***
ZonaDreta de l´Eixample   551.5946   108.3117   5.093 3.67e-07 ***
ZonaEsquerra Alta de l´Eixample   38.1628   117.2732   0.325 0.744879
ZonaEsquerra Baixa de l´Eixample   -33.9935   129.5046 -0.262 0.792956
ZonaFort Pienc   -90.1436   164.2074 -0.549 0.583058
ZonaGava Mar   414.2484   142.4937   2.907 0.003665 **
ZonaGlòries El Parc   92.4608   144.5754   0.640 0.522507
ZonaGòtic   -21.5107   115.4237 -0.186 0.852168
ZonaLes Corts   -63.3604   127.0067 -0.499 0.617891
ZonaPedralbes   545.1507   134.0476   4.067 4.84e-05 ***
ZonaPoble Sec   -135.7014   146.2146 -0.928 0.353405
ZonaPoblenou   -113.1630   143.3618 -0.789 0.429946
ZonaPutget - Farró   -154.7501   130.0786 -1.190 0.234238
ZonaRaval   -105.5832   115.8820 -0.911 0.362275
ZonaSagrada Família   -163.5453   133.5662 -1.224 0.220843
ZonaSant Antoni   -164.6999   140.4604 -1.173 0.241027
ZonaSant Gervasi - Bonanova   654.5794   119.3799   5.483 4.40e-08 ***
ZonaSant Gervasi - Galvany   314.8230   112.0311   2.810 0.004972 **
ZonaSant Ramon - Maternitat   -233.5699   146.3216 -1.596 0.110494
ZonaSants   -191.3438   144.2606 -1.326 0.184779
ZonaSarrià   677.9047   125.4398   5.404 6.83e-08 ***
ZonaSt. Pere - Sta. Caterina - El Born 152.3939   116.4582   1.309 0.190743
ZonaTres Torres   488.9765   145.5547   3.359 0.000787 ***
ZonaVallcarca - Penitents   -275.3804   167.7721 -1.641 0.100782
ZonaVila de Gràcia   -74.4534   125.8994 -0.591 0.554300
MoblatTRUE   -209.5491   29.0243 -7.220 6.04e-13 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 948.8 on 4713 degrees of freedom
Multiple R-squared: 0.6454, Adjusted R-squared: 0.6432
F-statistic: 286 on 30 and 4713 DF, p-value: < 2.2e-16
```

Veient els resultats, no totes les zones tenen un p.value prou baix com per a poder-nos fiar dels resultats. Per a les zones amb un p.value baix creiem que podríem fer servir les dades per a calcular possibles prediccions dels preus de lloguer.

Finalment provarem de crear un algoritme no supervisat de classificació, per a obtenir tres grups diferents d'apartaments.

```
```{r Probes_Estadistiques_3}
#3 Classificacio no supervisada
kMeansResult <- kmeans(dataset[,c("Sup_m2", "PreuActual")], centers=3)
kMeansResult$centers


cluster_dataset <- dataset
cluster_dataset$Cluster <- as.character(kMeansResult$cluster)

table(cluster_dataset$Cluster)
```
```

Resultats:
```

```
     Sup_m2 PreuActual
1 448.30702   9436.579
2 175.74413   3371.892
3  84.87845   1377.320

   1    2    3
 114  895 3735
```

Aquests tres grups els podríem arribar a anomenar, per exemple:

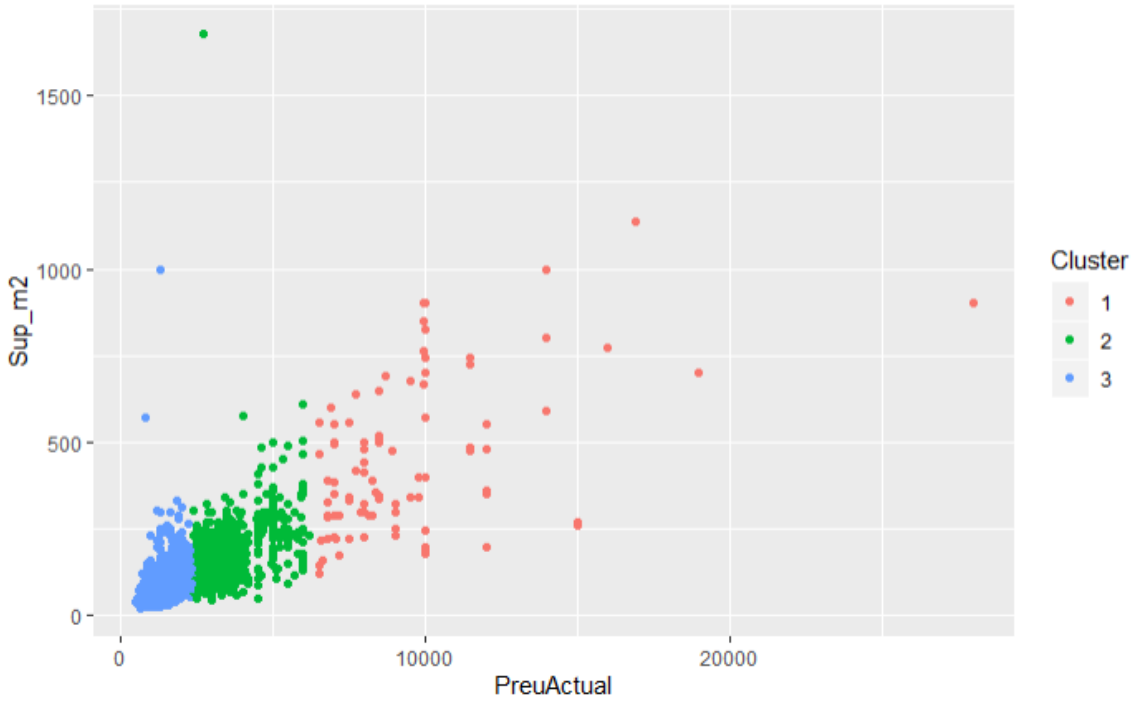
- Apartaments senzills: els que tenen el nucli a 84 m2 i amb un preu de 1377 euros al mes.
- Apartaments grans: amb el nucli a 175 m2 per 3371 euros al mes.
- Apartaments de luxe: amb el nucli als 448 m2 i en 9436 euros al mes. D'aquests , com podem veure, n'hi haurà més pocs.

Pregunta 5

Representació dels resultats a partir de taules i gràfiques

Gràfica amb els clústers trobats:

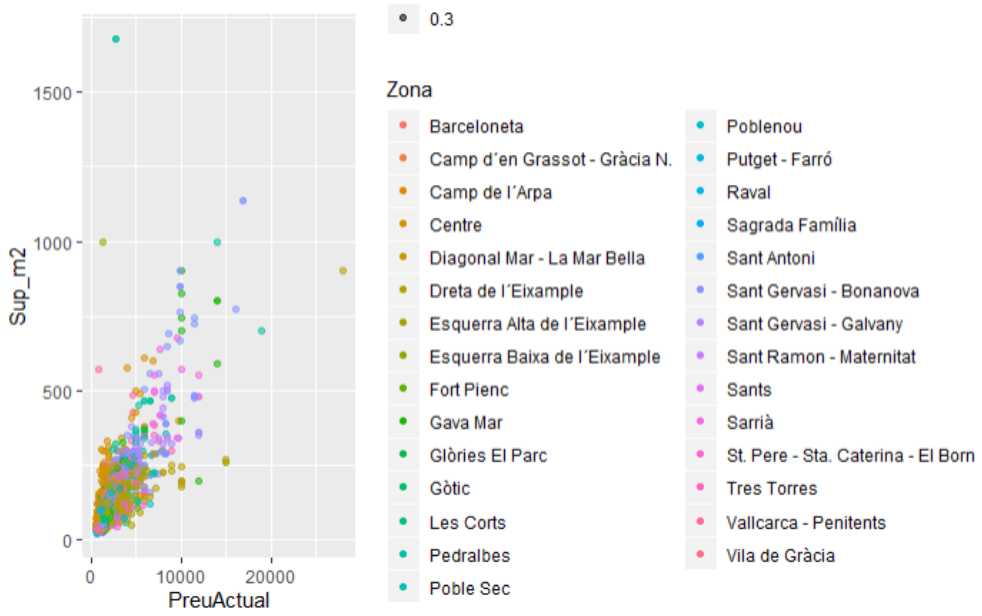
```
ggplot(data = cluster_dataset, mapping = aes(x = PreuActual, y = Sup_m2,
colour = Cluster)) + geom_point()
```



Veiem els tres grups d'habitatges trobats amb kmeans segons la superfície i el preu. S'intueix la relació lineal entre la superfície i el preu.

Gràfica relacionant preu i superfície segons cada zona:

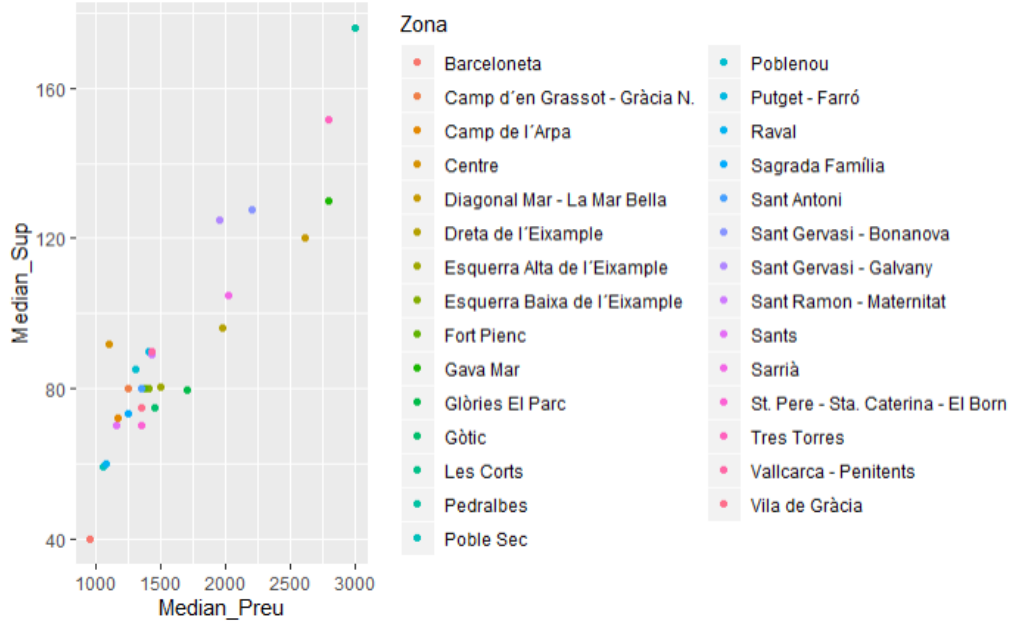
```
ggplot(data = dataset, mapping = aes(x = PreuActual, y = Sup_m2, color =
Zona, alpha = 0.3)) + geom_point()
```



La gràfica realment no dona informació molt directa si el que volem es comparar zones. Anem a veure què passa si treballem amb mesures centrals

Gràfica relacionant zones amb medianes de preu i superfície:

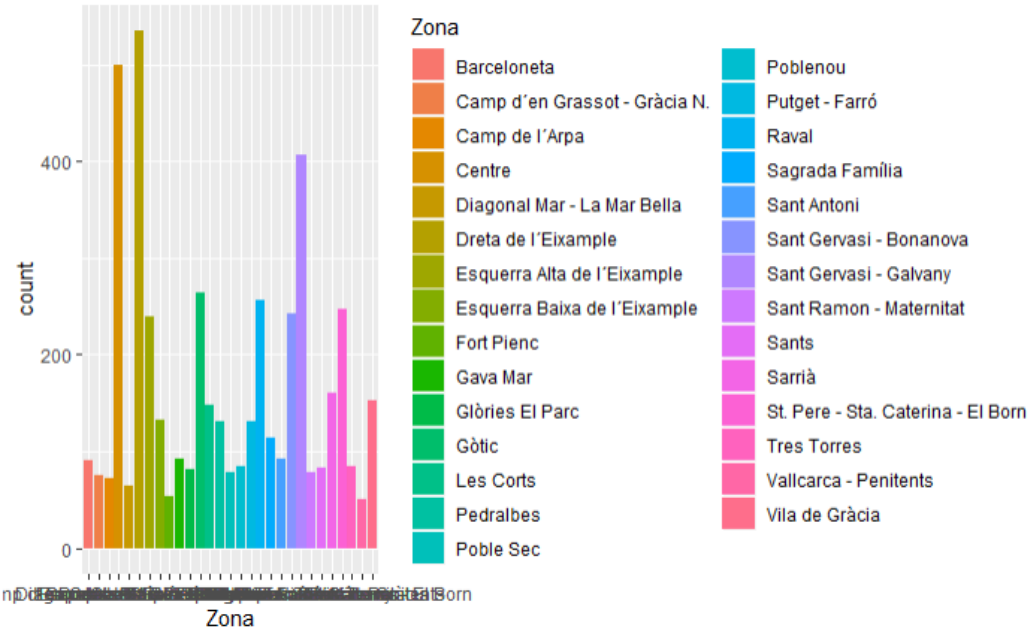
```
summarized_data <- dataset %>%
                    group_by(Zona) %>%
                      summarise(Median_Preu=median(PreuActual),
Median_Sup=median(Sup_m2))
ggplot(summarized_data, aes(x=Median_Preu, y=Median_Sup, colour=Zona)) +
geom_point()
```



Aquí sí que podem distingir amb més claredat quines zones tenen preus més alts i quines no, poden inferir quines són les zones més riques i quines són més humils.

Gràfica amb el nombre d'apartaments segons la zona:

```
ggplot(data = dataset, mapping = aes(x = Zona, fill = Zona)) + geom_bar()
```



Veiem que hi han 2 o 3 zones que destaquen per tenir molts apartaments en lloguer

Pregunta 6

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

Sí que som capaços de predir el els preus que poden tenir els lloguers definint la zona, la superfície de l'apartament i si estan moblats o no tal com es pot veure en el punt 4.3.2. El problema és que el resultat obtingut no es fiable per a totes les zones, nomes per a les que el p.value era menor que el valor de significació, típicament 0.05.

En general, doncs, donada una sèrie de dades sobre un habitatge de l'àrea de Barcelona, podríem fer una estimació prou acurada del seu preu de lloguer en una possible oferta. Com és d'esperar, la superfície en m2 és el millor predictor del preu final. Tot i que cal dir que la zona en què l'habitatge s'ubica és també molt rellevant quan comparem les zones amb preus més alts i més baixos.

Pregunta 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.

El codi font en R es pot trobar en el següent repositori de GitHub:

https://github.com/gikajavi/home-analytics

Taula de contribucions

Contribucions	Signatura
Investigació prèvia	MSS, JBL
Redacció de les respostes	MSS, JBL
Desenvolupament del codi	MSS, JBL