Web Scraping Project: www.habitaclia.com

Pol Ferrando & Wei Huang

May 23, 2017

Goal

The main goal of this project is to extract the important information about rental flats in Barcelona city: district, neighborhood, price, square meters, number of rooms and price per square meter.

With this information, we are going to compute the average price of rental flats in Barcelona by district and neighborhood, which could be useful either to know the price range in each zone or to choose the neighborhood or district in which look for according to the user's budget.

Data source

We will get the necessary data from Habitaclia's website[1], which is a portal that puts estate agents and individuals in touch with people looking to buy or rent a property.

This webpage is based on HTML, so we will scrap it using `rvest` R package and, after that, we will use regular expressions to clean the downloaded information in order to obtain the data of our interest.

Web scraping

We will scrap the webpage in two steps: a first one to extract the main information of each rental flat and a second one to obtain the district which a neighborhood belongs to.

First, we will extract the information of all rental ads in Barcelona city. One can found a maximum of 15 ads in each page, so we will need to scrap several urls. To do so, we have found a pattern in all these pages: "http://www.habitaclia.com/alquiler-barcelona-i.htm", where "i" is one of the search results pages (minus 1 because the first one does not match the pattern). Therefore, we can use a loop to scrap all these pages using the same code. However, there are some ads which do not include either the price or the description and we cannot use the same code to extract the same information, so we will treat them differently.[2]

```
require(rvest)
# number of results pages
url <- 'http://www.habitaclia.com/alquiler-barcelona.htm'
parsedURL <- read_html(url)
num <- parsedURL %>%
  html_nodes(".ultimo") %>%
  html_text()
num <- as.numeric(num)-1
# rental flats information
price <- NULL
info <- NULL
zone <- NULL
for(i in 0:num){
  # url
  if(i==0){
    url <- 'http://www.habitaclia.com/alquiler-barcelona.htm'
```

[1] www.habitaclia.com

[2] When this document was created, there were 339 search result pages for rental flats in Barcelona.

```
  }else{
    url <- paste0('http://www.habitaclia.com/alquiler-barcelona-',i,'.htm')
  }
  parsedURL <- read_html(url)
  # prices
  price.i <- parsedURL %>%
    html_nodes("li div span") %>%
    html_text()
  price.i <- grep("\u20AC", price.i, value=T)
  # info
  info.i <- parsedURL %>%
    html_nodes("div i") %>%
    html_text()
  # zone
  zone.i <- parsedURL %>%
    html_nodes("li div span div") %>%
    html_text()
  # if all ads of a page include all the information
  if(length(price.i)==15&length(info.i)==15&length(zone.i)==15){
    price <- c(price, price.i)
    info <- c(info, info.i)
    zone <- c(zone, zone.i)
  }
  # if any ad of a page does not include all the information
  else{
    indx <- max(grep("Barcelona", zone.i))
    # price
    price.i <- parsedURL %>%
      html_nodes(".opciones") %>%
      html_text()
    price.i <- price.i[grep("Avísame", price.i)]
    price.i <- substr(price.i, 5, regexpr("\u20AC", price.i))
    price <- c(price, price.i[1:indx])
    # info
    info2 <- parsedURL %>%
      html_nodes(".datos") %>%
      html_text()
    info.i <- NULL
    for(j in 1:length(info2)){
      aux <- strsplit(info2[j], "\t\t\t\t")[[1]][6]
      info.i <- c(info.i, ifelse(grepl("-", aux), aux, NA))
    }
    info <- c(info, info.i[1:indx])
    # zone
    zone <- c(zone, zone.i[1:indx])
  }
}
head(price, 3)
```

```
## [1] "2.400 <U+0080>" "2.000 <U+0080>" "1.800 <U+0080>"
```

```
head(info, 3)
```

```
## [1] "94m2 - 2 habitaciones - 25,53 <U+0080> /m2"
```

```
## [2] "105m2 - 3 habitaciones - 19,05 <U+0080> /m2"
## [3] "102m2 - 3 habitaciones - 17,65 <U+0080> /m2"
```

```
head(zone, 3)
```

```
## [1] "Barcelona Tres Torres\r\n"    "Barcelona Putget - Farró\r\n"
## [3] "Barcelona Sant Antoni\r\n"
```

Therefore, we need to clean these data, which can be done using regular expressions:

```
price2 <- as.numeric(gsub("\\s\u20AC|\\.", "", price))
m2 <- as.numeric(gsub("\\.","",substr(info,1,regexpr("m2",info)-1)))
rooms <- as.numeric(substr(info,regexpr("[0-9]+\\shabitaci",info),regexpr("\\shabitaci",info)-1))
pricem2 <- as.numeric(gsub(",",".",substr(info,regexpr("[0-9]+,[0-9]+\\s\u0080",info),
                                          regexpr("\\s\u0080",info)-1)))
neighborhood <- gsub("\r\n", "", zone)
neighborhood <- gsub("Barcelona\\s", "", neighborhood)
bcn.data <- data.frame(neighborhood, price=price2, rooms, m2, pricem2)
head(bcn.data)
```

```
##                         neighborhood price rooms  m2 pricem2
## 1                       Tres Torres  2400     2  94   25.53
## 2                     Putget - Farró  2000     3 105   19.05
## 3                       Sant Antoni  1800     3 102   17.65
## 4 St. Pere - Sta. Caterina - El Born   850     1  50   17.00
## 5            Sant Gervasi - Bonanova  1590     3  90   17.67
## 6                          Poblenou  1600     4 126   12.70
```

Second, we will extract the district of each neighborhood of Barcelona[3]:

```
url <- 'http://www.habitaclia.com/alquiler-vivienda-en-barcelona/provincia_barcelona-barcelones-area_6/l
parsedURL <- read_html(url)
district <- parsedURL %>%
  html_nodes(".verticalul") %>%
  html_text()
district <- strsplit(district, "\\s+[0-9]+\\.?[0-9]+(\r\n)+")[[1]]
#install.packages("stringi", dependencies=TRUE)
library(stringi)
district2 <- tolower(district)
district2 <- gsub("\\s", "_", district2)
district2 <- stri_trans_general(district2,"Latin-ASCII")
distr.nb <- NULL
for(i in 1:length(district2)){
  url <- paste0("http://www.habitaclia.com/alquiler-vivienda-en-barcelona-distrito_",district2[i],
                "/provincia_barcelona-barcelones-area_6/seldistrito.htm")
  parsedURL <- read_html(url)
  aux <- parsedURL %>%
    html_nodes(".verticalul") %>%
    html_text()
  aux <- strsplit(aux, "\\s*[0-9]*\\.?[0-9]*(\r\n)+")[[1]]
  distr.nb <- rbind(distr.nb, data.frame(district=district[i], neighborhood=aux, stringsAsFactors=F))
}
head(distr.nb)
```

```
##       district                      neighborhood
## 1 Ciutat Vella                       Barceloneta
```

[3] http://www.habitaclia.com/alquiler-vivienda-en-barcelona/provincia_barcelona-barcelones-area_6/buscardistrito.htm

```
## 2 Ciutat Vella                                Gòtic
## 3 Ciutat Vella                                Raval
## 4 Ciutat Vella St. Pere - Sta. Caterina - El Born
## 5     Eixample                Dreta de l´Eixample
## 6     Eixample        Esquerra Alta de l´Eixample
```

And merge this information with the previous data to obtain our dataset for the analysis:

```
bcn.data.final <- merge(distr.nb, bcn.data, by="neighborhood", all.y=T)
head(bcn.data.final)
```

```
##   neighborhood       district price rooms  m2 pricem2
## 1        Badal Sants Montjuïc  1000     3  70   14.29
## 2        Badal Sants Montjuïc  1350     3 120   11.25
## 3        Badal Sants Montjuïc   885     3  52   17.02
## 4        Badal Sants Montjuïc   900     3  70   12.86
## 5        Badal Sants Montjuïc  1922     3  63   30.51
## 6        Badal Sants Montjuïc  1350     3  70   19.29
```

Analysis

The summary of the data is:

```
summary(bcn.data.final)
```

```
##  neighborhood         district            price             rooms
##  Length:5099        Length:5099        Min.   :    370   Min.   : 1.000
##  Class :character   Class :character   1st Qu.:   1100   1st Qu.: 2.000
##  Mode  :character   Mode  :character   Median :   1500   Median : 3.000
##                                        Mean   :   2492   Mean   : 2.637
##                                        3rd Qu.:   2200   3rd Qu.: 3.000
##                                        Max.   :2595000   Max.   :20.000
##                                        NA's   :8         NA's   :86
##        m2            pricem2
##  Min.   :   6.0   Min.   :  1.00
##  1st Qu.:  65.0   1st Qu.: 14.19
##  Median :  85.0   Median : 17.27
##  Mean   : 107.4   Mean   : 19.30
##  3rd Qu.: 120.0   3rd Qu.: 21.73
##  Max.   :5309.0   Max.   :228.41
##  NA's   :13       NA's   :20
```

Note that there are some missing values and, also, some strange prices, square meters and prices per square meter (we have checked that they correspond to errors in the description). Therefore, we remove these observations from the data:

```
bcn.data.final <- with(bcn.data.final, bcn.data.final[!is.na(price)&!is.na(m2)&!is.na(pricem2)&
                                                       price!=2595000&m2>10&pricem2>1,])
summary(bcn.data.final)
```

```
##  neighborhood         district            price           rooms
##  Length:5075        Length:5075        Min.   :  370   Min.   : 1.000
##  Class :character   Class :character   1st Qu.: 1100   1st Qu.: 2.000
##  Mode  :character   Mode  :character   Median : 1500   Median : 3.000
##                                        Mean   : 1984   Mean   : 2.634
##                                        3rd Qu.: 2200   3rd Qu.: 3.000
```

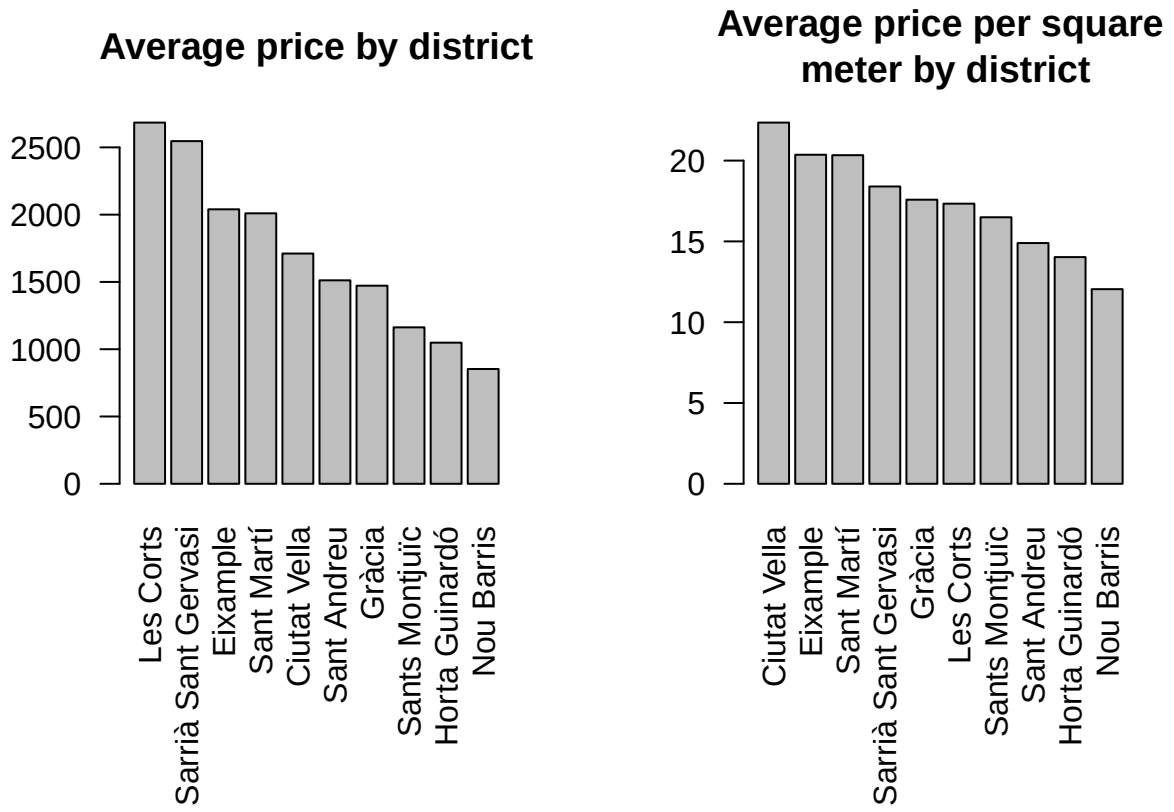
```
##                                         Max.   :45000   Max.   :20.000
##                                                        NA's   :82
##        m2            pricem2
##  Min.   :  20.0   Min.   :  5.17
##  1st Qu.:  65.0   1st Qu.: 14.19
##  Median :  85.0   Median : 17.27
##  Mean   : 105.8   Mean   : 19.20
##  3rd Qu.: 120.0   3rd Qu.: 21.71
##  Max.   :1577.0   Max.   :137.78
##
```

Now, we can compute and plot both the average price and the average price per square meter by district:

```
distr.avg <- aggregate(cbind(price,pricem2) ~ district, data=bcn.data.final, mean)
(distr.avg <- distr.avg[order(distr.avg$price, decreasing=T),])
```

```
##               district     price  pricem2
## 5            Les Corts 2684.0451 17.33546
## 10 Sarrià Sant Gervasi 2546.4509 18.39875
## 2              Eixample 2039.7262 20.36269
## 8            Sant Martí 2009.9759 20.33532
## 1          Ciutat Vella 1711.0926 22.35184
## 7           Sant Andreu 1511.9239 14.89902
## 3               Gràcia 1472.0868 17.58251
## 9        Sants Montjuïc 1162.6681 16.49447
## 4        Horta Guinardó 1048.7376 14.02807
## 6            Nou Barris  853.1429 12.04467
```

```
par(mfrow=c(1,2), mar=c(9,4,4,2))
barplot(distr.avg$price, names.arg=distr.avg$district, las=2, cex.names=1,
        main="Average price by district")
distr.avg <- distr.avg[order(distr.avg$pricem2, decreasing=T),]
barplot(distr.avg$pricem2, names.arg=distr.avg$district, las=2, cex.names=1,
        main="Average price per square\n meter by district")
```



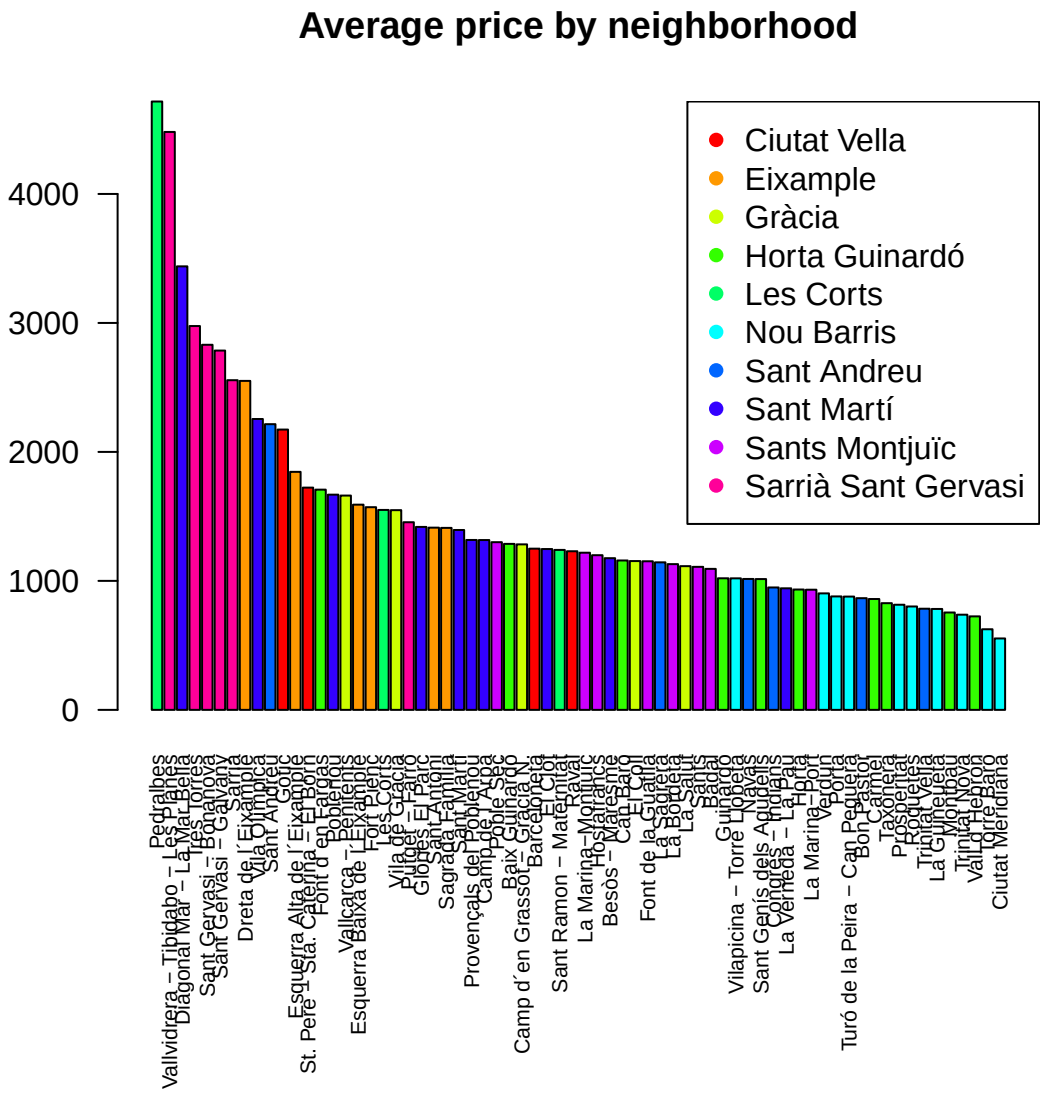
On the left plot, we can see that the most expensive districts are Les Corts and Sarrià Sant Gervasi, with an average price approximately three times larger than the one of the cheapest districts (Horta Guinardó and Nou Barris). However, on the right plot we can see that Les Corts and Sarrià Sant Gervasi are not the most expensive districts, and they have medium prices per square meter. Instead, Ciutat Vella has the highest price per square meter, which is probably because there are many small flats with high prices.

Finally, we can compute and plot the average price by neighborhood:

```
nb.pr <- aggregate(cbind(price,pricem2) ~ neighborhood + district, data=bcn.data.final, mean)
nb.pr <- nb.pr[order(nb.pr$price, decreasing=T),]
head(nb.pr)
```

```
##                           neighborhood           district    price
## 27                           Pedralbes          Les Corts 4716.692
## 68 Vallvidrera - Tibidabo - Les Planes Sarrià Sant Gervasi 4480.800
## 47         Diagonal Mar - La Mar Bella          Sant Martí 3438.774
## 67                          Tres Torres Sarrià Sant Gervasi 2975.904
## 64              Sant Gervasi - Bonanova Sarrià Sant Gervasi 2830.702
## 65               Sant Gervasi - Galvany Sarrià Sant Gervasi 2785.734
##     pricem2
## 27 19.43135
## 68 12.85767
## 47 28.92365
## 67 17.61771
## 64 20.18836
## 65 18.58746
```

```
palette(rainbow(10))
par(mar=c(10,4,4,2))
barplot(nb.pr$price, names.arg=nb.pr$neighborhood, col=factor(nb.pr$district),
        las=2, cex.names=0.7, main="Average price by neighborhood")
legend("topright", levels(factor(nb.pr$district)),
       col=1:length(levels(factor(nb.pr$district))), pch=16)
```

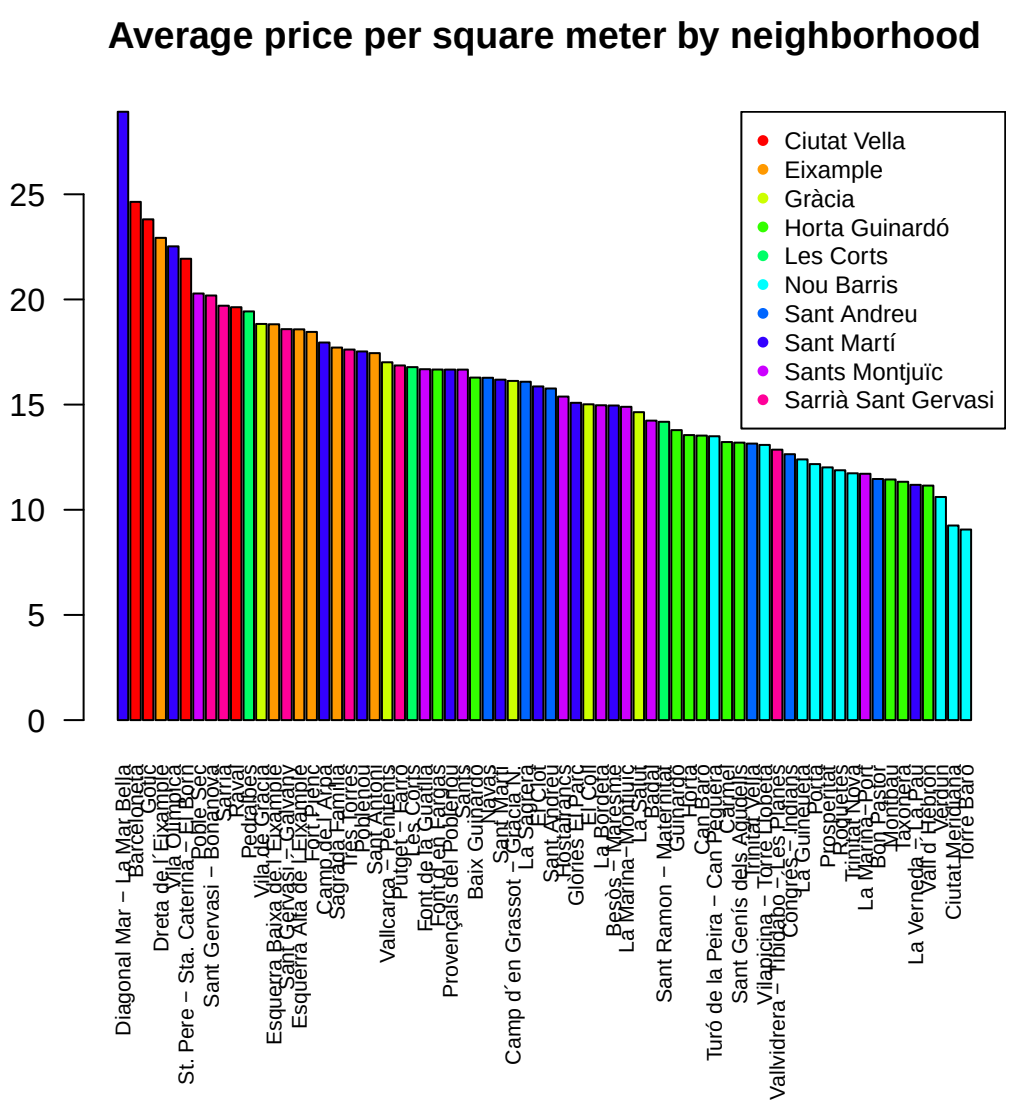


Firstly, now it is clear that Les Corts district had the higher average prices because Pedralbes is the most expensive neighborhood in Barcelona, but the average prices in Les Corts and Sant Ramon - Maternitat (the other two neighborhoods that belong to Les Corts district) are quite standard. Secondly, note that all neighborhoods in Sarrià Sant Gervasi except Putget Farró are very expensive in average. Finally, we can see that the cheapest flats (in average) are located in neighborhoods of Nou Barris and Horta Guinardó, which explains why these two districts had the lowest average prices.

Also, we can compute and plot the average price per square meter by neighborhood:

```
par(mar=c(10,4,4,2))
nb.pr <- nb.pr[order(nb.pr$pricem2, decreasing=T),]
```

```
barplot(nb.pr$pricem2, names.arg=nb.pr$neighborhood, col=factor(nb.pr$district),
        las=2, cex.names=0.7, main="Average price per square meter by neighborhood")
legend("topright", levels(factor(nb.pr$district)),
       col=1:length(levels(factor(nb.pr$district))), pch=16, cex=0.75)
```



Again, we can conclude that, in Ciutat Vella, the prices are too high in relation to the area of the flats, which leads to the highest prices per square meter. Also, note that Diagonal Mar - La Mar Bella (that belongs to Sant Martí) has the highest value, while the rest of neighborhoods of this district are quite low.

Conclusions

With this project, we have learned that a basic knowledge of HTML, regular expressions and the R package `rvest` is a powerful tool for data analysts to scrap the web and get the necessary data for analysis.

In our case, a limitation that we have found during the scraping process is that there were some pages containing ads which did not have all the required information for our analysis, so we had to change our scraping code in these cases in order to overcome this issue and get all the data we wanted.

In summary, scraping Habitaclia's website has given us the chance to collect information about rental flats in Barcelona and we have used it to have a first idea of the average prices in every neighborhood and district of the city, but note that we could use the obtained dataset to do more sophisticated analysis. Moreover, for a complete analysis of all rental flats in Barcelona, we could have scraped more websites similar to Habitaclia's (such as *idealista* or $fotocasa$) because with this project we only consider the ads posted in www.habitaclia.com.