Trabajo práctico 2: Estadística Descriptiva

2022-09-09

El dataset

Para este notebook emplearemos el dataset de Properati, que contiene información de propiedades del 2019.

En la siguiente celda lo cargamos:

```
datos_full <- read.csv('ar_properties.csv', stringsAsFactors = FALSE)
```

Ahora construimos un dataset más pequeño con las propiedades de Boedo, Colegiales, Centro / Microcentro, Mataderos y Puerto Madero.

```
datos_CABA <- datos_full[datos_full$l2 == 'Capital Federal',]
barrios <- c("Boedo", "Colegiales", "Centro / Microcentro", "Mataderos", "Puerto Madero")
datos_barrios <- datos_CABA[is.element(datos_CABA$l3, barrios), ]
```

La siguiente tabla muestra la cantidad de propiedades en cada barrio:

```
tabla_barrios <- table(datos_barrios$l3)
print(tabla_barrios)
```

```
##
##                Boedo Centro / Microcentro           Colegiales
##                 1155                 1848                 1844
##            Mataderos        Puerto Madero
##                 1185                 2463
```

Como existen datos faltantes en términos de superficie total, superficie cubierta y número de ambientes, procedemos a eliminar las propiedades que los contengan.

```
indices <- !is.na(datos_barrios$rooms) & !is.na(datos_barrios$surface_covered) & !is.na(datos_barrios$surface_tot
al) & !is.na(datos_barrios$l3) & !is.na(datos_barrios$price)

datos_barrios <- datos_barrios[indices, ]
```

Tamaño de las propiedades

Para nuestro análisis vamos a usar las variables:

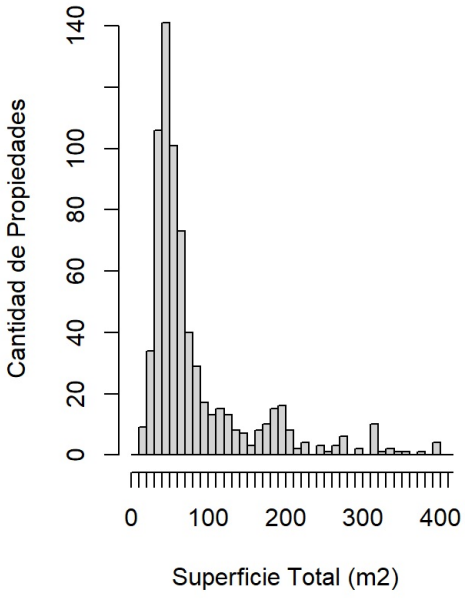
```
- surface_total: La superficie total de la propiedad.
- rooms: El número de habitaciones que tiene la propiedad.
```

Primero vamos a visualizar la superficie total y el número de habitaciones típicas en cada barrio.

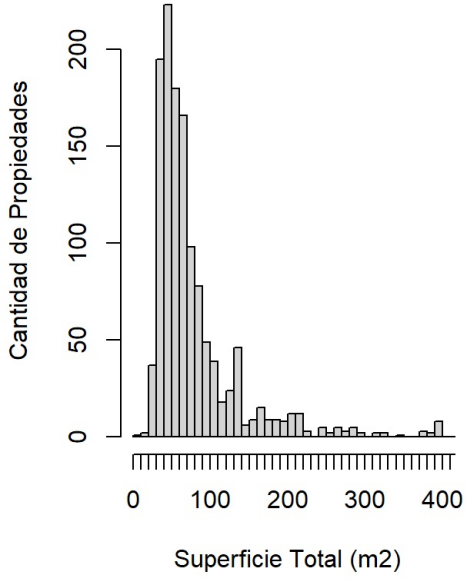
Procedemos a hacer un histograma para mostrar la cantidad de propiedades en función de su superficie total en cada barrio.

```
# Ordenamos los graficos en una matriz de 2 columnas
par(mfrow=c(1, 2))

# Usamos un for loop para automatizar el proceso
for (i in barrios) {
  props <- datos_barrios[datos_barrios$l3 == i,]
  superficie_maxima <- max(props$surface_total)
  max_rango <- quantile(props$surface_total, 0.99)
  breakss <- seq(0,superficie_maxima+10,10)

# Le pedimos que nos muestre en el eje horizontal el intervalo en el que caen el 99% de los valores de superficie
(por barrio), ya que si no, hay datos muy aislados que distraen la atencion del grueso de los datos.
# Elegimos el tamaño de los bins "a ojo", luego de observar en qué orden se encuentra el porcentaje considerado d
el dataset. Así, en este caso, obtenemos también la misma escala en el eje horizontal para cada barrio, facilitan
do la
# comparación entre ellos.

  hist(props$surface_total, breaks = breakss, xlab = 'Superficie Total (m2)', ylab = 'Cantidad de Propiedades', m
ain = i, xlim
       = c(0,max_rango))
  axis(side = 1, at = breakss, label=NA)
}
```

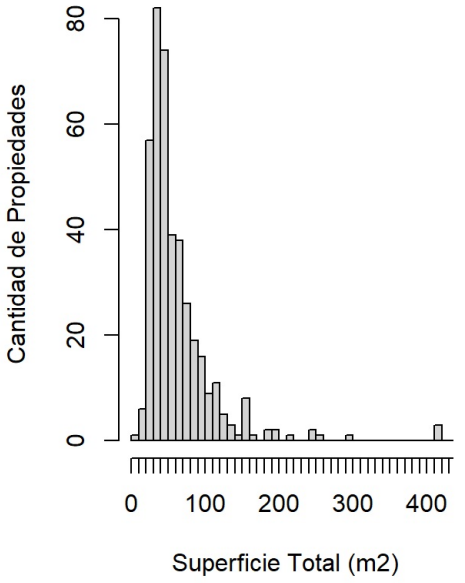
Boedo



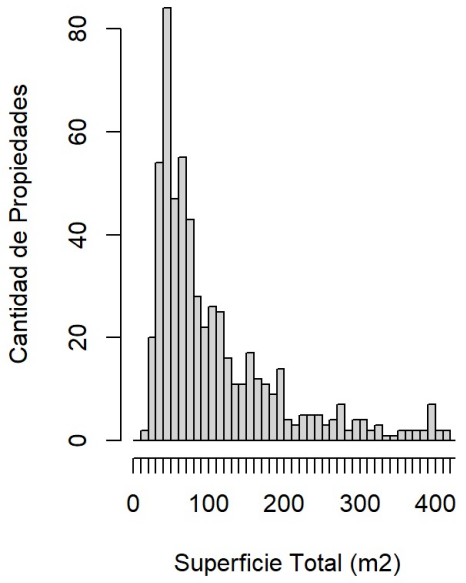
Colegiales



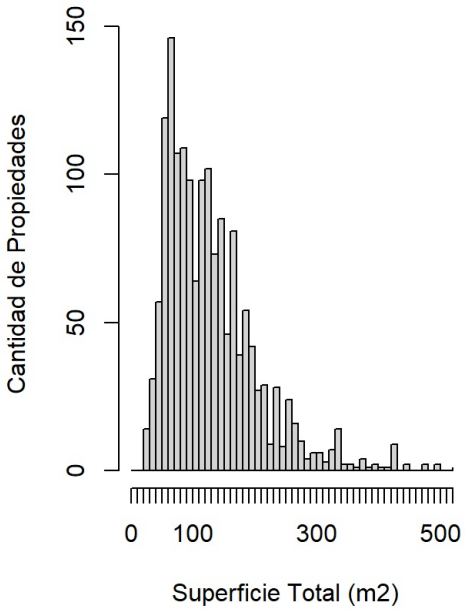
Centro / Microcentro



Mataderos



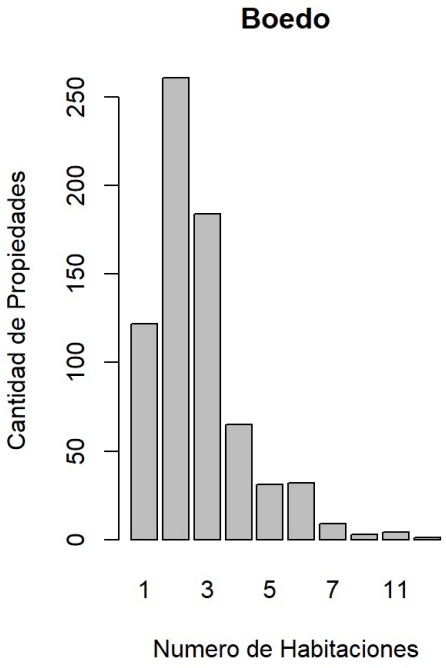
Puerto Madero

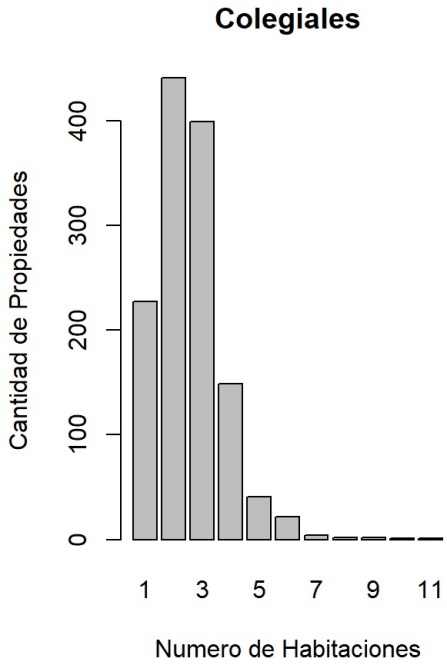


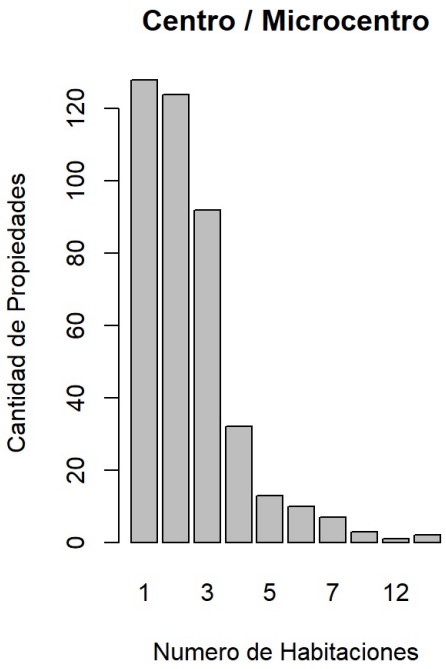
En estos gráficos se observa que la superficie total típica en general se centra entre los 30 y los 50 metros cuadrados aproximadamente, salvo en Puerto Madero donde hay una cantidad de propiedades significativa entre los 50 y 200 metros cuadrados, es decifr, con una variación mayor en superficie total del grueso de las propiedades en comparación con el resto.

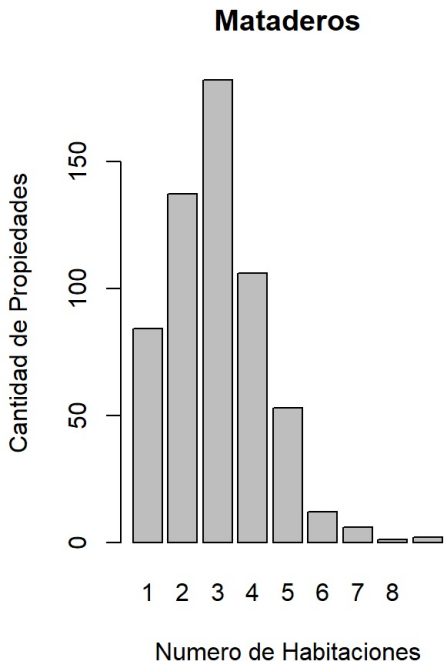
Para visualizar el número de habitaciones de cada barrio, hacemos un barplot para cada uno.

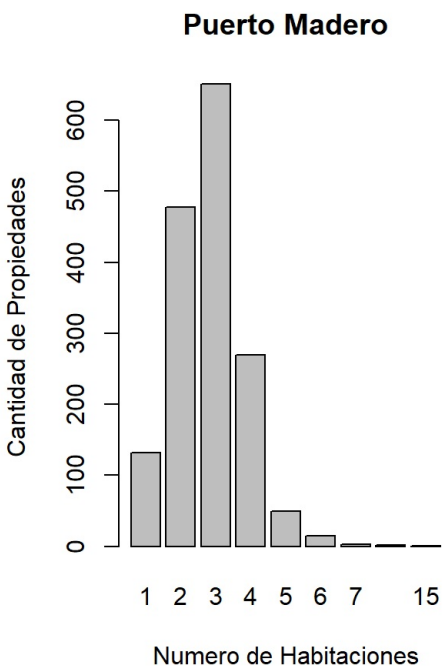
```
par(mfrow=c(1, 2))

for (i in barrios) {
  props <- datos_barrios[datos_barrios$l3 == i,]
  barplot(table(props$rooms), main = i, xlab = 'Numero de Habitaciones', ylab = 'Cantidad de Propiedades')
}
```


Boedo
Cantidad de Propiedades
0
50
100
150
200
250
1
3
5
7
11
Numero de Habitaciones


Colegiales
Cantidad de Propiedades
0
100
200
300
400
1
3
5
7
9
11
Numero de Habitaciones


Centro / Microcentro
Cantidad de Propiedades
0
20
40
60
80
100
120
1
3
5
7
12
Numero de Habitaciones


Mataderos
Cantidad de Propiedades
0
50
100
150
1
2
3
4
5
6
7
8
Numero de Habitaciones


Puerto Madero
Cantidad de Propiedades
0
100
200
300
400
500
600
1
2
3
4
5
6
7
15
Numero de Habitaciones

Es fácil ver que en Centro/Microcentro tiende a haber más monoambientes que cualquier otro tipo de propiedad (aunque también hay casi la misma cantidad de propiedades con 2 ambientes), a comparación de otros barrios estudiados más residenciales, como Boedo o Colegiales por ejemplo, donde la proporción de monoambientes es menor que la de propiedades con más de una habitación.

Ahora vamos a caractizar la superficie total de una propiedad típica en cada barrio. Para esto podemos mirar las medidas sumarias de la variable en cuestión (surface_total).

```
#Imprimimos las medidas sumarias de cada variable
for(i in barrios) {
  props <- datos_barrios[datos_barrios$l3 == i,]
  supTotal <- props$surface_total
  print(i)
  print(summary(supTotal))
}
```

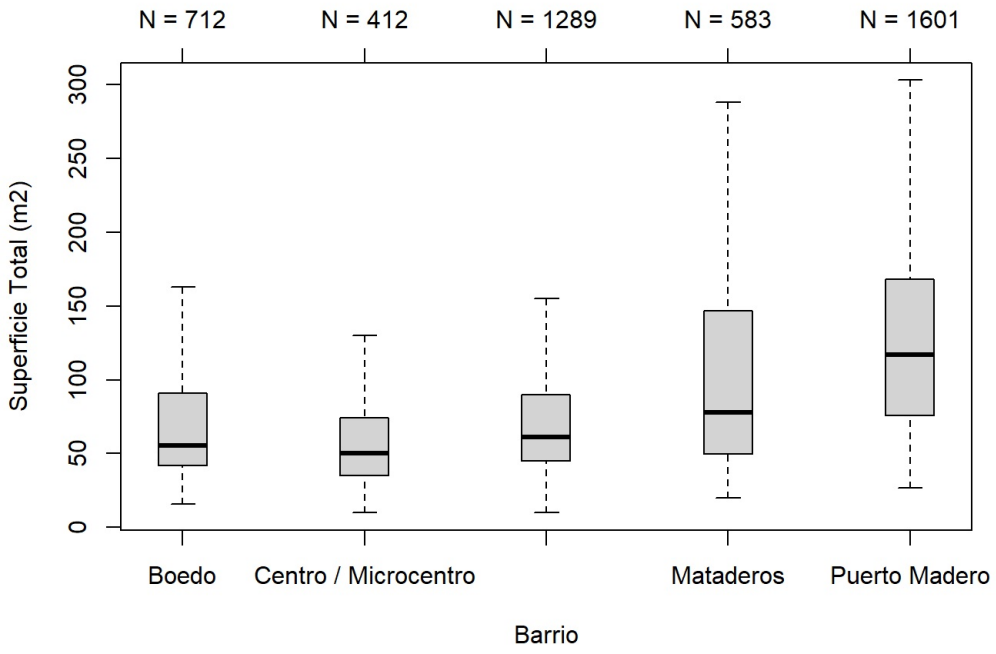
```
## [1] "Boedo"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   16.00   42.00   55.50   92.25   91.00 4421.00
## [1] "Colegiales"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.00   45.00   61.00   83.59   90.00  715.00
## [1] "Centro / Microcentro"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.00   35.00   50.00   89.17   74.00 4415.00
## [1] "Mataderos"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    20.0    49.5    78.0   113.2   147.0   560.0
## [1] "Puerto Madero"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      27      76     117     140     168    7971
```

Como podemos observar en la tabla, la media siempre es mayor que la mediana. Sin embargo, estas magnitudes se hallan en órdenes de magnitud próximos, en comparación con el rango de valores que toma la superficie total de una propiedad en cada barrio. Esto se debe a que la media esta siendo influenciada por valores altos poco representativos, lo cual trae como consecuencia una distribución asimétrica de los datos, provocando el corrimiento de estos hacia izquierda y por ello la media se ubica más “hacia la derecha” que la mediana.

En general, las superficies varían entre los 10 m^2 y los 7971 m^2 aunque el rango no es el mismo en todos los barrios. El tamaño de una propiedad alcanza la máxima variabilidad en Puerto Madero, donde se haya la propiedad con la superficie más grande (7971 m^2), mientras que el mínimo rango se da en Mataderos, y la propiedad con menor superficie se debe de hallar en Colegiales o Centro/Microcentro.

Esta información se puede visualizar de forma resumida y clara en los siguientes boxplots:

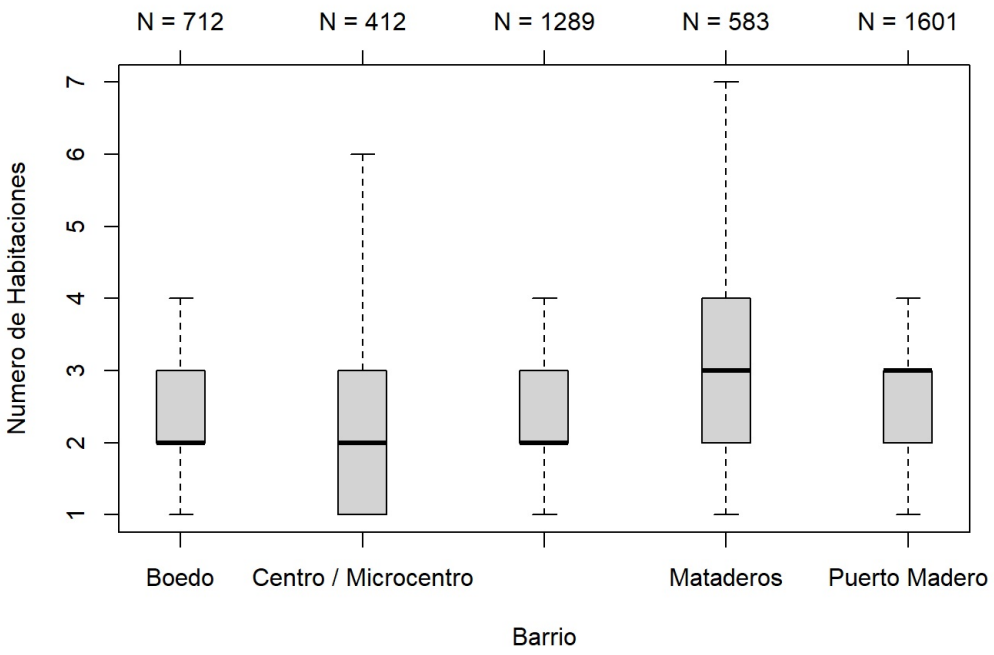
```
boxplot(datos_barrios$surface_total ~ datos_barrios$l3, outline = FALSE, xlab = "Barrio", ylab = "Superficie Tota
l (m2)",
        at = seq(1,5 * 3, by = 3))
axis(side = 3, at = seq(1, 5 * 3, by = 3), label = paste('N =', table(datos_barrios$l3)))
```



Se puede observar que el rango de superficie total que nos muestra el boxplot para cada barrio es significativamente menor que el rango de la tabla. Esto se debe a que en el gráfico quitamos los outliers de modo de quedarnos con los datos representativos. No es útil este tipo de gráfico para responder los puntos anteriores, pero nos permite observar más directamente otro tipo de cuestiones. Por ejemplo, podemos ver que las medianas, los cuantiles y los bigotes de las superficies de Boedo y Colegiales son muy similares, por lo que las chances de encontrar en estos barrios propiedades con igual superficie deben ser idénticas.

Podemos proceder análogamente para resumir el número de habitaciones que suelen tener las propiedades en cada barrio.

```
boxplot(datos_barrios$rooms ~ datos_barrios$l3, outline = FALSE, xlab = "Barrio", ylab = "Numero de Habitaciones"
,
        at = seq(1,5 * 3, by = 3))
axis(side = 3, at = seq(1,5 * 3, by = 3), label = paste('N =', table(datos_barrios$l3)))
```



Como habiamos visto anteriormente, hay una presencia importante de monoambientes en Centro/Microcentro (se incluyen también las oficinas y cocheras). Tan es así que es el único barrio donde las propiedades con 1 habitación están dentro de la caja, es decir, se encuentran dentro del 50% central del total junto con las habitaciones de 2 y 3 ambientes.

Por lo demás no vemos nada fuera de lo usual, las propiedades de 2 y 3 ambientes son las mas populares y en Mataderos notamos que también hay una cantidad significativa de propiedades con 4 ambientes.

```
#Creamos la variable habProm que mide la superficie que tiene la habitacion promedio de cada propiedad.
datos_barrios[,"habProm"] <- datos_barrios$surface_covered / datos_barrios$rooms

#Utilizamos la funcion mean para sacar el promedio de habProm en cada barrio: esto es la superficie que tiene una
habitación promedio en todo el barrio.
for(i in barrios) {
  props <- datos_barrios[datos_barrios$l3 == i, ]
  promedio <- mean(props$habProm)
  print(paste(i, ":"))
  print(promedio)
}
```

```
## [1] "Boedo :"
## [1] 27.71468
## [1] "Colegiales :"
## [1] 27.86108
## [1] "Centro / Microcentro :"
## [1] 36.095
## [1] "Mataderos :"
## [1] 30.12878
## [1] "Puerto Madero :"
## [1] 43.46252
```

Dentro de los barrios que estamos estudiando, en Puerto Madero hay habitaciones más grandes, mientras que en Boedo y Colegiales se hallan habitaciones mas pequeñas.

Relación entre el precio y las características de una propiedad

En esta parte, estudiaremos las variables: - surface_covered: La superficie cubierta (techada) de la propiedad. - price: El precio de una propiedad. - property_type: El tipo de una propiedad. - operation_type: El tipo de operación disponible para una propiedad.

Retenemos las propiedades con precios en dólares:

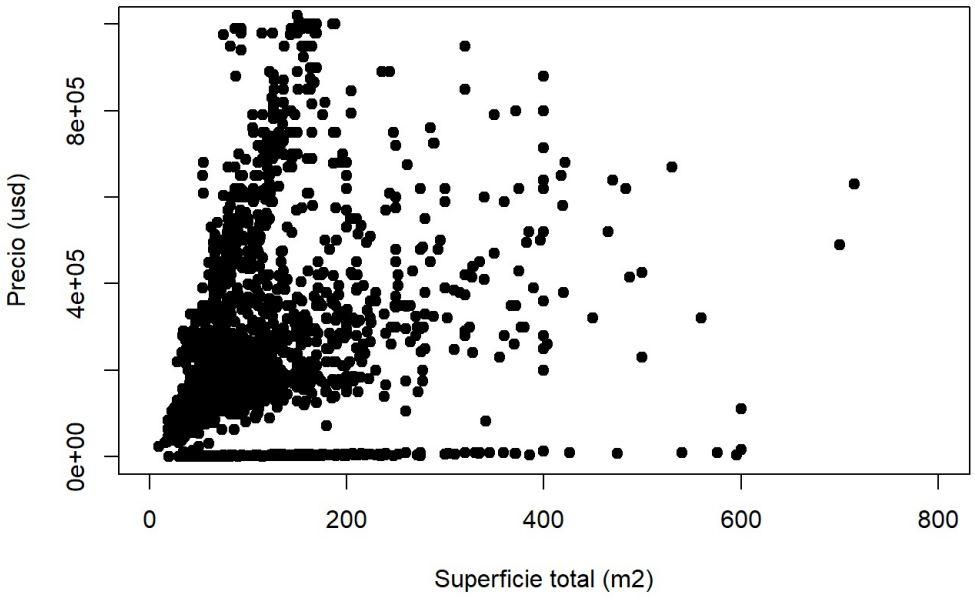
```
datos_usd <- datos_barrios[datos_barrios$currency == 'USD',]
```

Agregamos una columna que muestre el fondo (superficie total - superficie cubierta):

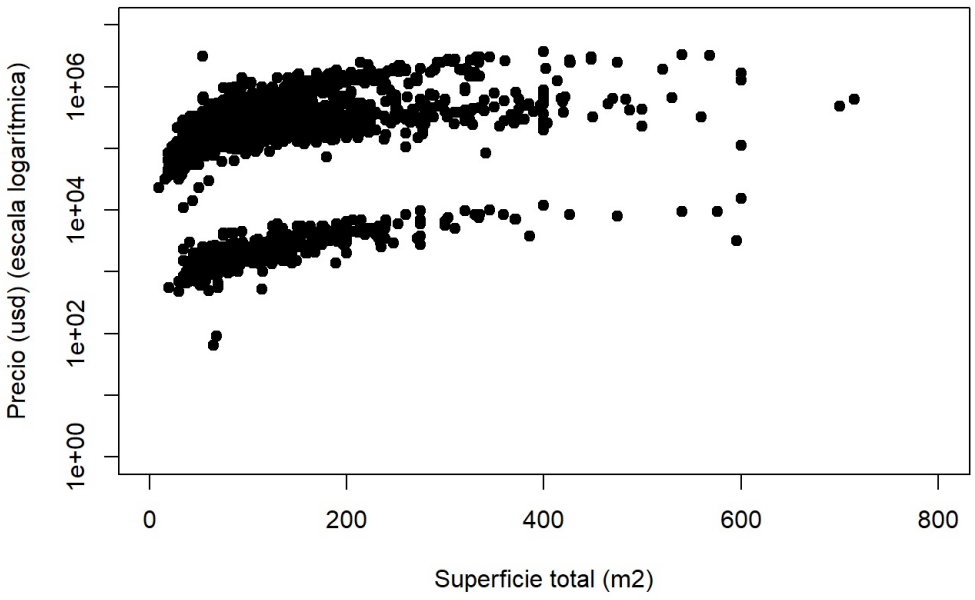
```
datos_usd$fondo <- datos_usd$surface_total - datos_usd$surface_covered
```

Graficamos el precio de la propiedad en función de la superficie total y cubierta, utilizando dos escalas: una lineal y otra logarítmica. Esta última nos es útil dado que los datos en el eje y son mucho mayores a los datos en el eje x.

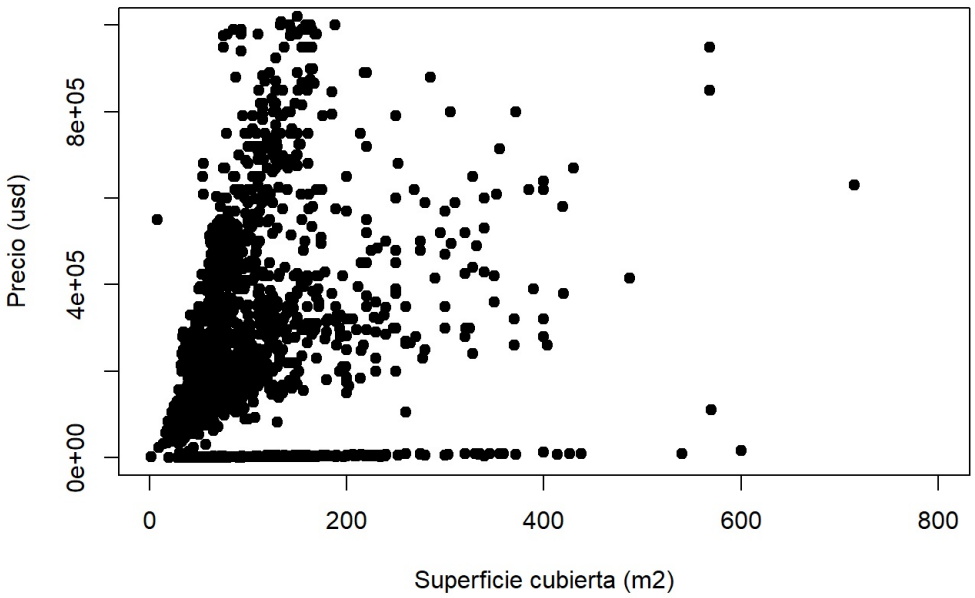
```
#Sup. total
plot(datos_usd$surface_total,
     datos_usd$price,
     xlim=c(1,8e2),ylim=c(1,1e6),xlab='Superficie total (m2)',ylab='Precio (usd)',pch=1
)
points(datos_usd$surface_total,
     datos_usd$price,
     xlim=c(1,8e2),ylim=c(1,1e6),xlab='Superficie total (m2)',ylab='Precio (usd)',pch=16
)
```



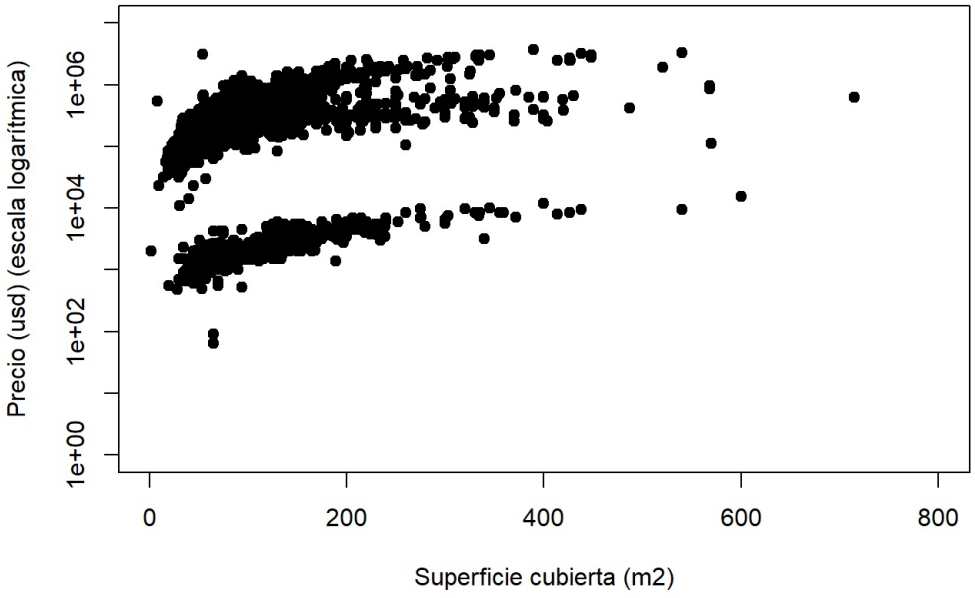

```
plot(datos_usd$surface_total,
     datos_usd$price,
     xlim=c(1,8e2),ylim=c(1,1e7),xlab='Superficie total (m2)',ylab='Precio (usd) (escala logarítmica)',pch=1, log
= "y"
)
points(datos_usd$surface_total,
     datos_usd$price,
     xlim=c(1,8e2),ylim=c(1,1e7),xlab='Superficie total (m2)',ylab='Precio (usd)',pch=16
)
```




```
#Sup. cubierta
plot(datos_usd$surface_covered,
     datos_usd$price,
     xlim=c(1,8e2),ylim=c(1,1e6),xlab='Superficie cubierta (m2)',ylab='Precio (usd)',pch=1,
)
points(datos_usd$surface_covered,
     datos_usd$price,
     xlim=c(1,8e2),ylim=c(1,1e6),xlab='Superficie cubierta (m2)',ylab='Precio (usd)',pch=16
)
```



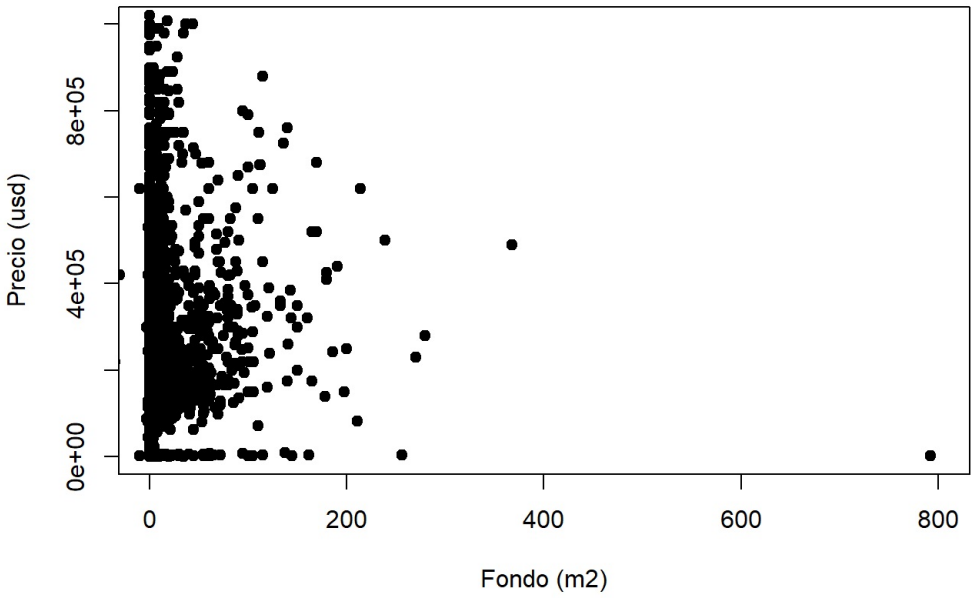

```
plot(datos_usd$surface_covered,
     datos_usd$price,
     xlim=c(1,8e2),ylim=c(1,1e7),xlab='Superficie cubierta (m2)',ylab='Precio (usd) (escala logarítmica)',pch=1,
log = "y"
)
points(datos_usd$surface_covered,
     datos_usd$price,
     xlim=c(1,8e2),ylim=c(1,1e7),xlab='Superficie cubierta (m2)',ylab='Precio (usd)',pch=16
)
```



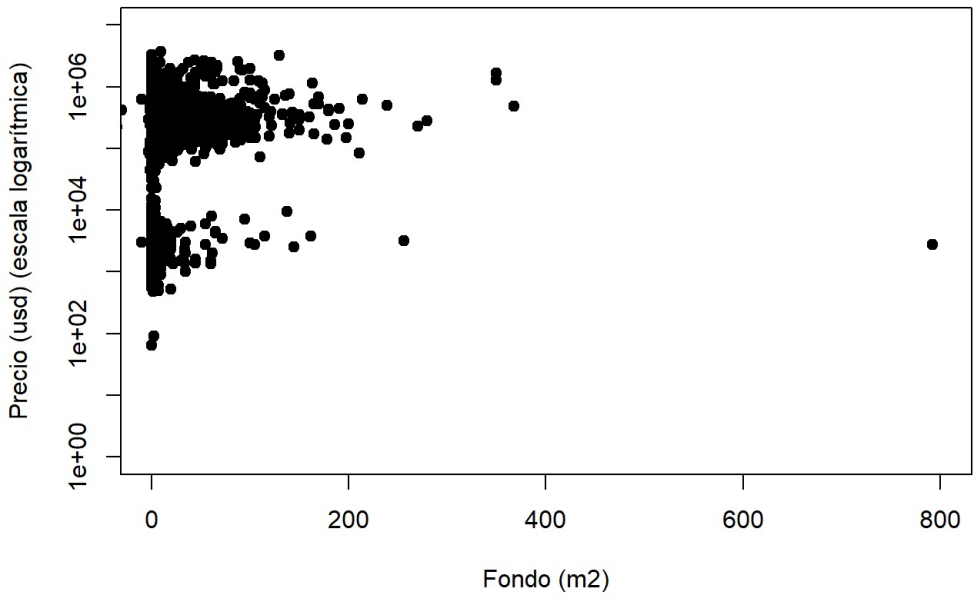
Los gráficos de precio en función de superficie (tanto cubierta como total) muestran que, en general, a mayor superficie, mayor es el precio. De hecho, ambos gráficos son muy similares, de manera que podemos concluir que la superficie cubierta se comporta como la superficie total.

Ahora graficamos el precio de la propiedad en función del fondo y del tamaño promedio de habitación, utilizando nuevamente dos escalas diferentes.

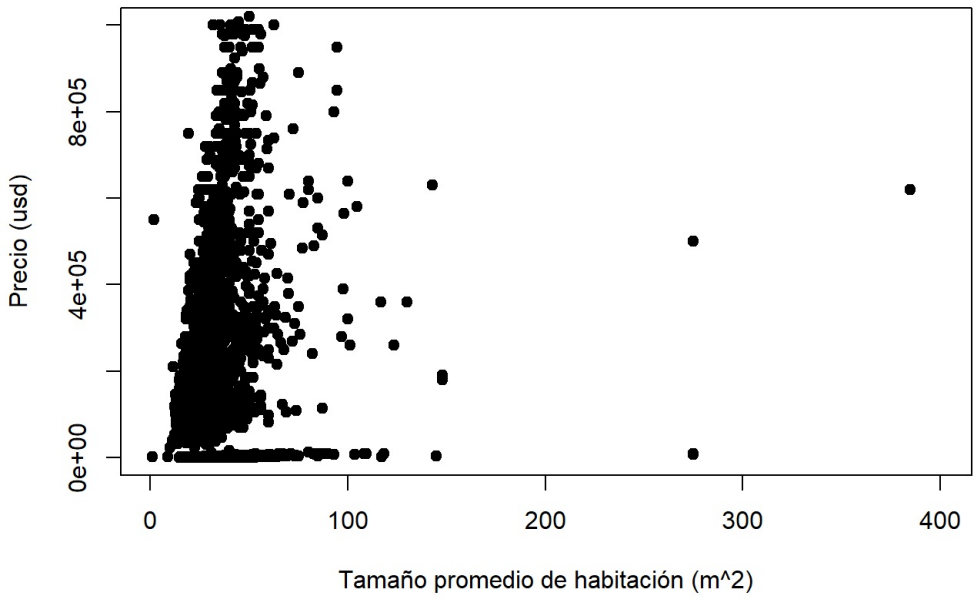
```
#Fondo
plot(datos_usd$fondo,
     datos_usd$price,
     xlim=c(1,8e2),ylim=c(1,1e6),xlab='Fondo (m2)',ylab='Precio (usd)',pch=1,
)
points(datos_usd$fondo,
     datos_usd$price,
     xlim=c(1,8e2),ylim=c(1,1e6),xlab='Fondo (m^2)',ylab='Precio (usd)',pch=16
)
```



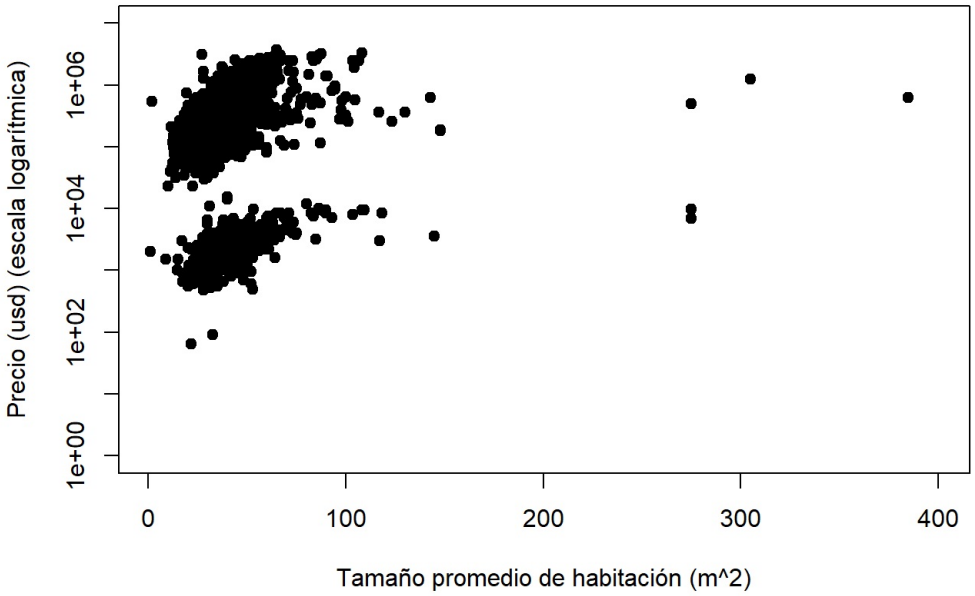
```
plot(datos_usd$fondo,
     datos_usd$price,
     xlim=c(1,8e2),ylim=c(1,1e7),xlab='Fondo (m2)',ylab='Precio (usd) (escala logarítmica)',pch=1, log = "y"
)
points(datos_usd$fondo,
     datos_usd$price,
     xlim=c(1,8e2),ylim=c(1,1e7),xlab='Fondo (m^2)',ylab='Precio (usd)',pch=16
)
```




```
#Hab. prom
plot(datos_usd$habProm,
     datos_usd$price,
     xlim=c(1,4e2),ylim=c(1,1e6),xlab='Tamaño promedio de habitación (m^2)', ylab = 'Precio (usd)', pch=1
)
points(datos_usd$habProm,
     datos_usd$price,
     xlim=c(1,4e2),ylim=c(1,1e6),xlab='Tamaño promedio de habitación (m^2)',ylab='Precio (usd)',pch=16
)
```



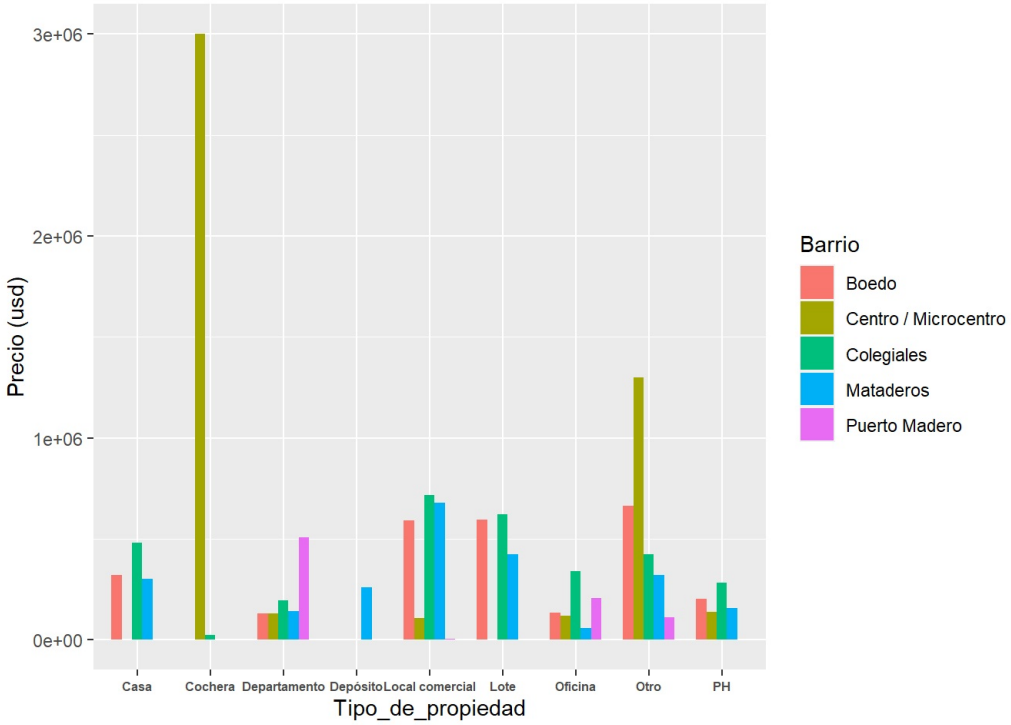
```
plot(datos_usd$habProm,
     datos_usd$price,
     xlim=c(1,4e2),ylim=c(1,1e7),xlab='Tamaño promedio de habitación (m^2)', ylab = 'Precio (usd) (escala logarít
mica)', pch=1, log = "y"
)
points(datos_usd$habProm,
     datos_usd$price,
     xlim=c(1,4e2),ylim=c(1,1e7),xlab='Tamaño promedio de habitación (m^2)',ylab='Precio (usd)',pch=16
)
```



Los gráficos muestran que no necesariamente un mayor fondo o tamaño de habitación aumentan el precio. La superficie cubierta, en cambio, es un factor mucho más relevante a la hora de cotizar una propiedad. El precio de una habitación promedio, depende de la superficie cubierta: a mayor tamaño promedio, mayor es el precio. Por esto, volvemos a ver un comportamiento lineal en el gráfico correspondiente, pero con una pendiente más pronunciada que en el gráfico de precio en función de superficie cubierta.

Ahora exploraremos el precio en función del tipo de propiedad, y su distribución en cada barrio.

```
# Creamos un vector con los tipos de propiedad
prop_types <- levels(factor(datos_usd$property_type))

# Limpio el dataset de datos con precio faltante
datos_usd <- datos_usd[!is.na(datos_usd$price), ]

# Creamos un nuevo mini dataset con el promedio de precios de los barrios por tipo de propiedad
cantidad_filas_dataset <- length(prop_types) * length(barrios)
datos_promedio <- data.frame(
            'Barrio' = rep('-', cantidad_filas_dataset),
            'Tipo_de_propiedad' = rep('-', cantidad_filas_dataset),
            'Precio' = rep(0, cantidad_filas_dataset)
)

contador <- 1
for (i in barrios) {
  for (j in prop_types) {
    datos_promedio$Barrio[contador] <- i
    datos_promedio$Tipo_de_propiedad[contador] <- j
    precio_promedio <- datos_usd$price[datos_usd$l3 == i & datos_usd$property_type == j]
    datos_promedio$Precio[contador] <- mean(precio_promedio)
    contador <- contador + 1
  }
}

datos_promedio$Precio[is.nan(datos_promedio$Precio)] <- 0

# Usamos ggplot para hacer un gráfico de barras agrupadas
library(ggplot2)

ggplot(datos_promedio, aes(x = Tipo_de_propiedad, y = Precio, fill = Barrio)) + geom_col(width = 0.7, position="d
odge") + theme(axis.text.x = element_text(size=6,face="bold")) + ylab("Precio (usd)")
```




```
table(datos_usd$operation_type)
```

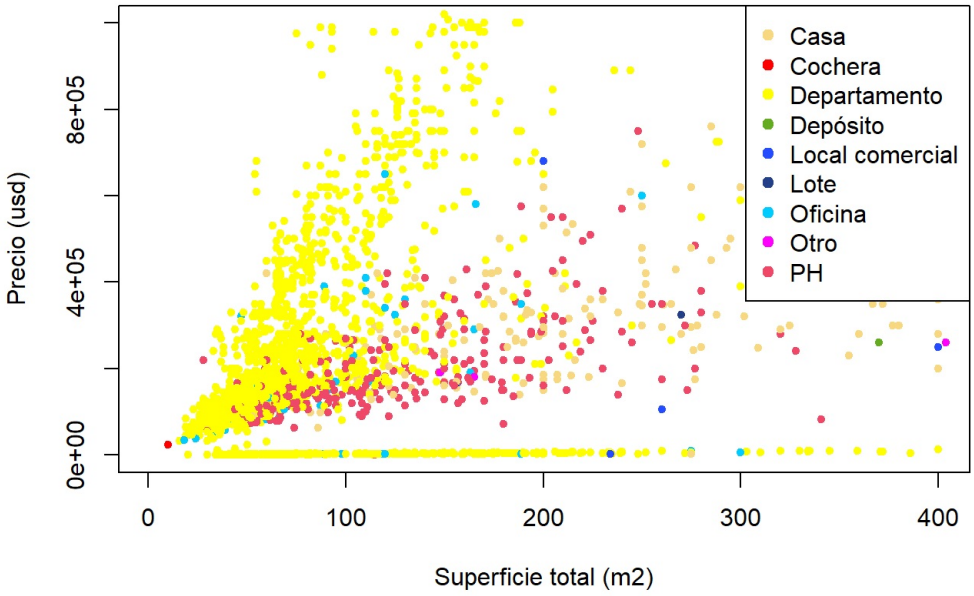
```
##
##          Alquiler Alquiler temporal             Venta
##               443               172              2888
```

Observamos que no en todos los barrios se encuentran los mismos tipos de propiedades, por ejemplo no hay casas en Puerto Madero, ni hay lotes en Centro / Microcentro. Además, el precio de un mismo de tipo de propiedad, varía según el barrio en que esta se halla: podemos ver que las casas en Colegiales suelen ser más caras que en Boedo o en Mataderos, o que los departamentos suelen ser más caros en Puerto Madero que en cualquier otro barrio.

Hay que tener en cuenta que en este gráfico se están considerando simultaneamente operaciones en venta como en alquiler: las propiedades en venta aportan significativamente más al promedio que aquellas que están en alquiler, puesto que los precios de estas últimas son más bajos que si estuviesen en venta. No obstante, viendo la cantidad de propiedades en venta y en alquiler, solamente el 25% del total se corresponde con estas últimas. Puede verse que el gráfico no cambia significativamente al estudiar solamente las propiedades en venta.

En la siguiente celda graficamos un scatter plot del precio en función de la superficie total, además diferenciamos cada tipo de propiedad (departamento, casa, oficina, etc.) con un color distinto.

```
#eleccion de colores
colors <- c("#F9D882","#FF0000","#FFFF00", "#65AC22", "#264CFF", "#244089","#00CCFF","#FF00FF","#EF4868", "#CC00F
F")


plot(datos_usd$surface_total,
     datos_usd$price, xlim=c(1,4e2) , ylim=c(1,1e6),
     col = colors[factor(datos_usd$property_type)],
     xlab='Superficie total (m2)',ylab='Precio (usd)',pch=20
)

legend('topright',legend = prop_types,
       pch = 19,
       col = colors)
```



Podemos corroborar nuevamente que a mayor superficie, mayor es el precio. Igualmente, el precio depende también del tipo de propiedad, pues por ejemplo, hay casas con una superficie mayor a 200m^2 que valen menos que un departamento de 100m^2.

En general, en la oferta de propiedades predominan los departamentos, cuyas superficies no varían tanto como las casas y PHs. Esto se ve en la concentración de puntos correspondientes a departamentos hacia la izquierda del gráfico, en contraste con la dispersión de puntos que corresponden a casas y PHs hacia la derecha.

Podemos ver una línea inferior horizontal donde se concentran mayormente puntos asociados a departamentos con precios considerablemente bajos en relación al resto de propiedades. Creemos que esto se debe a que el tipo de operación corresponde a un alquiler en vez de una venta.

Conclusiones finales

En general, las propiedades que ofrece Properati tienen entre 2 y 3 habitaciones, con una presencia predominante de monoambientes en Microcentro por ejemplo. No obstante, no ocurre que en todos los barrios se hallan los mismos tipos de propiedades, viendo que no encontramos casas en Puerto Madero, aunque la oferta de departamentos es superior a la de cualquier otra propiedad en general.

El precio de las propiedades depende en parte del tamaño que estas tengan: en este punto observamos que comparar superficie total con superficie cubierta resulta casi equivalente, pues esto es lo que cobra más relevancia a la hora de cotizar una propiedad. De todas maneras, esta no es la única variable que lo define, sino que también influye por ejemplo el tipo de propiedad en cuestión, como el barrio en que esta se encuentra: los departamentos en Puerto Madero son mucho más caros que en Colegiales, o una oficina de 100m^2 no sale lo mismo que una casa con igual superficie.