extracción de datos

Diego Castillo

13/11/2020

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(mice)
```

```
##
## Attaching package: 'mice'

## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```
library(VIM)
```

```
## Loading required package: colorspace

## Loading required package: grid

## Loading required package: data.table

##
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
## VIM is ready to use.
##  Since version 4.0.0 the GUI is in its own package VIMGUI.
##
##           Please use the package to use the new (and old) GUI.

## Suggestions and bug-reports can be submitted at: https://github.com/alexkowa/VIM/issues

##
## Attaching package: 'VIM'

## The following object is masked from 'package:datasets':
##
##     sleep
```

```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

Introducción

En el presente proyecto se realizará un análisis con el fin de encontrar las relaciones más importantes entre los atributos del mercado del alquiler de España y así resolver las preguntas de cuales son los patrones más importantes que tiene este mercado así como los sesgos del mismo. Para el desarrollo del proyecto de explotarán los datos de la Encuesta de Condiciones de Vida (INE, 2019), así como la información disponible del portal inmobiliario Idealista.

La Encuesta de Condiciones de Vida (ECV) es una operación estadística que se lleva a cabo de manera anual por el INE, cuyo objetivo es la producción sitemática de estadísticas comunitarias sobre la renta y las condiciones de vida de la población. Incluye datos transversales y longitudinales comparables y actualizados sobre la renta, el nivel y composición de la pobreza y la exclusión social, a escala nacional y europea.

Entre los diferentes aspectos a los cuales la ECV está orientada, es de especial relevancia para el presente proyecto los referentes a los ingresos y situación económica de los hogares, empleo, vivienda, movimientos migratorios y nivel de formación.

El portal inmobiliario Idealista es uno de los principales sitios web de anuncios de alquiler y venta de inmuebles, funciona en varios países europero y tiene vasta presencia en España desde hace ya varios años. Idealista pone a disposición una API pública mediante la cual se puede realizar la consulta de la información del alquiler y venta según el lugar deseado.

En primer lugar, se realiza la extracción, tratamiento y exploración de los datos de la ECV.

Extracción de la información

Se realiza la extracción de la información correspondiente a la Encuesta de Condiciones de Vida (ECV) obtenida desde la web oficial del INE para el periodo 2019. El INE pone a disposición del público en general los datos de la ECV, los cuales están conformados por 4 ficheros en formato csv los cuales contienen información básica y detallada de los hogares y personas encuestadas. En el presente proyecto se realizará el análisis de los datos detallados de hogares y personas.

```
hogares <- read.csv('../data/data_collect/enc_condiciones_vida/esudb19h.csv', header= TRUE, sep= ',', s
personas <- read.csv('../data/data_collect/enc_condiciones_vida/esudb19p.csv', header= TRUE, sep= ',', s
```

Data cleaning

Conjunto de datos de *Hogares*

El proceso de limpieza de la información consta de la eliminación de atributos que no son relevantes para el análisis que se realizará, más bien son atributos que cuentan con información descriptiva de la encuesta en si. Las variables a eliminar se listan a continuación:

Variable	Descripción
HB010	Año de la encuesta
HB020	País
HB050	Mes de la entrevista
HB050_F	Mes de la entrevista - Flag
HB060	Año de la entrevista
HB060_F	Año de la entrevista - Flag
HB070	Identificación personal del informante del cuestionario sobre el hogar
HB070_F	Identificación personal del informante del cuestionario sobre el hogar - Flag
HB080	Identificación de la primera persona responsable de la vivienda
HB080_F	Identificación de la primera persona responsable de la vivienda - Flag
HB090	Identificación de la segunda persona responsable de la vivienda
HB090_F	Identificación de la segunda persona responsable de la vivienda - Flag
HB100	Numero de minutos empleados en cumplimentar el cuestionario sobre el hogar
HB100_F	Numero de minutos empleados en cumplimentar el cuestionario sobre el hogar - Flag

```
# Se genera un vector con las variables se eliminarán del data frame
col_borrar_hogares <- c('HB010','HB020','HB050','HB050_F','HB060','HB060_F','HB070','HB070_F','HB080','H

# Se eliminan las variables innecesarias
hogares <-hogares %>% select(-one_of(col_borrar_hogares))
```

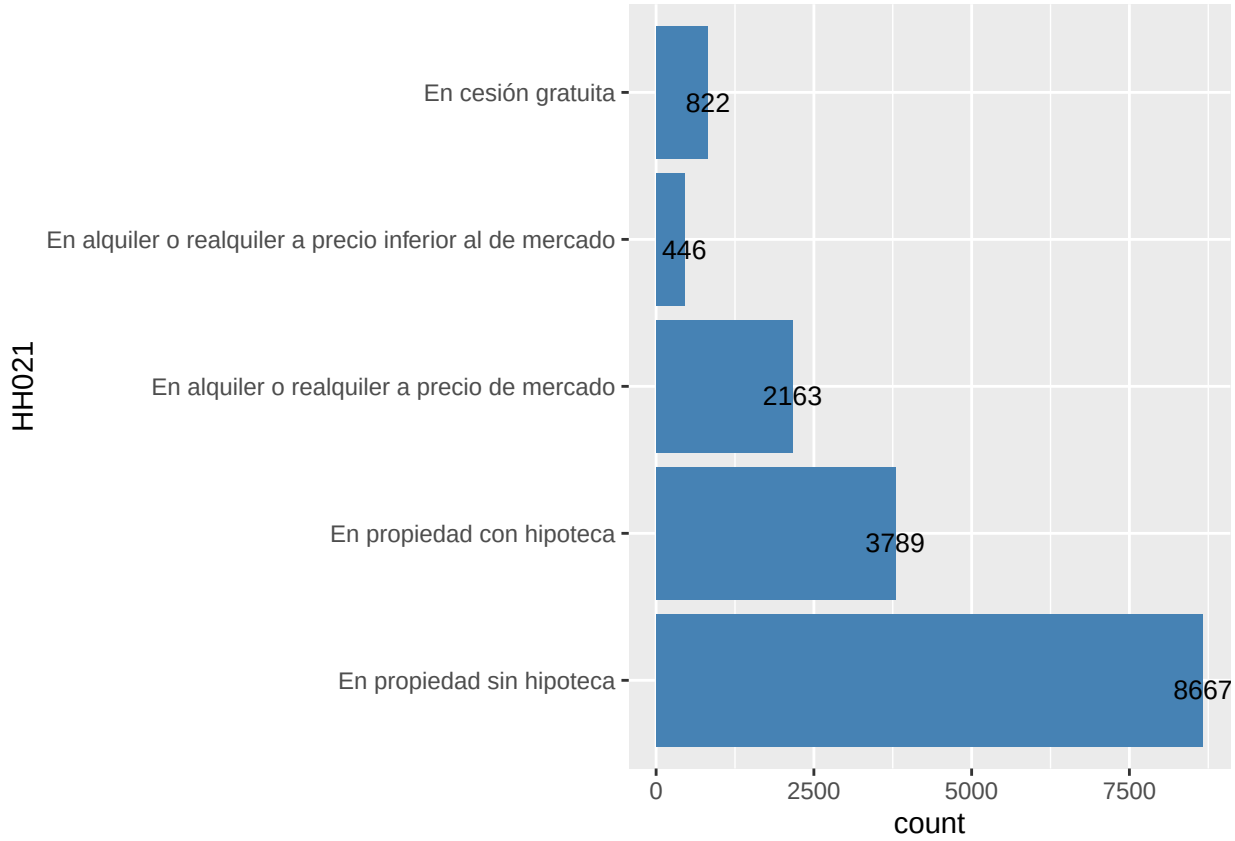
Análisis exploratorio de datos

Régimen de tenencia

La variable `HH021` de la ECV almacena el régimen de tenencia del hogar encuestado, por lo que primero se analiza como se distribuye la información dentro de esta variable.

```
# Se coloca los labels para ver el gráfico de manera correcta.
hogares$HH021 <- factor(hogares$HH021,levels = c(1,2,3,4,5),labels = c('En propiedad sin hipoteca','En p
```

```r
# se realiza el gráfico
g1.hh021 <- ggplot(hogares, aes(HH021)) +
  geom_bar(fill="steelblue") +
  geom_text(aes(label=..count..), stat='count', vjust=1, color="black", size=3.5) +
  coord_flip()
g1.hh021
```



El régimen de tenencia predominante es el de la propiedad con un total de `12.456` hogares encuestados que corresponde al `78.4%` del total de la población de la encuesta. El motivo del presente análisis son aquellas viviendas que están en un régimen de alquiler, en donde se tiene que la tenencia de alquiler a precio de mercado se cuenta con un total de `2.163` registros que corresponde al `13,61%` y para el régimen de tenencia de alquiler a precio inferior al mercado se tienen `446` registros que equivale al `2,8%`.

El objetivo del presente proyecto es realizar el análisis de los hogares y las personas que residen en esos hogares en un régimen de tenencia de alquiler. No se toma en cuenta para el análisis los registro que corresponden a los hogares con un régimen de tenencia de alquiler a precio inferior al mercado, ya que dichos precios pueden afectar el valor real del mercado del alquiler.

```r
hogares.alquiler <- filter(hogares, HH021 == 'En alquiler o realquiler a precio de mercado')
```

Precio del alquiler

La variable con más relevancia del conjunto de datos, es aquella que contiene el valor que las familias pagan por el alquiler de una vivienda en la que residen. Dentro del conjunto de datos, esta variable se denota como

HH060. Se analiza en primera instancia si dentro de los hogares con tenencia en alquiler se tienen valores nulos.

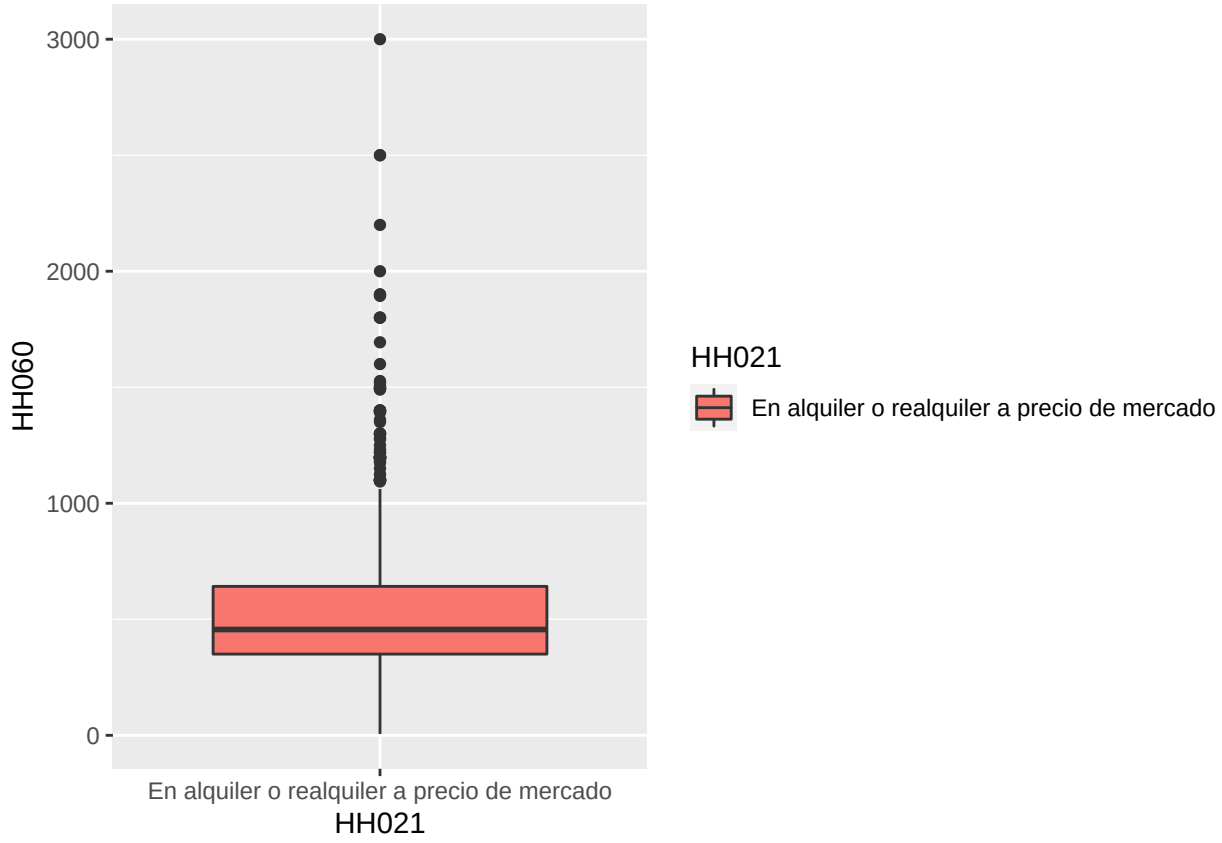
```
sum(is.na(hogares.alquiler$HH060))
```

```
## [1] 32
```

Existen 32 valores vacios o nulos para la variable HH060, se aplicará en este caso la imputación por ...

Se analiza la distribución del precio del alquiler de las viviendas utilizando un diagrama de cajas.

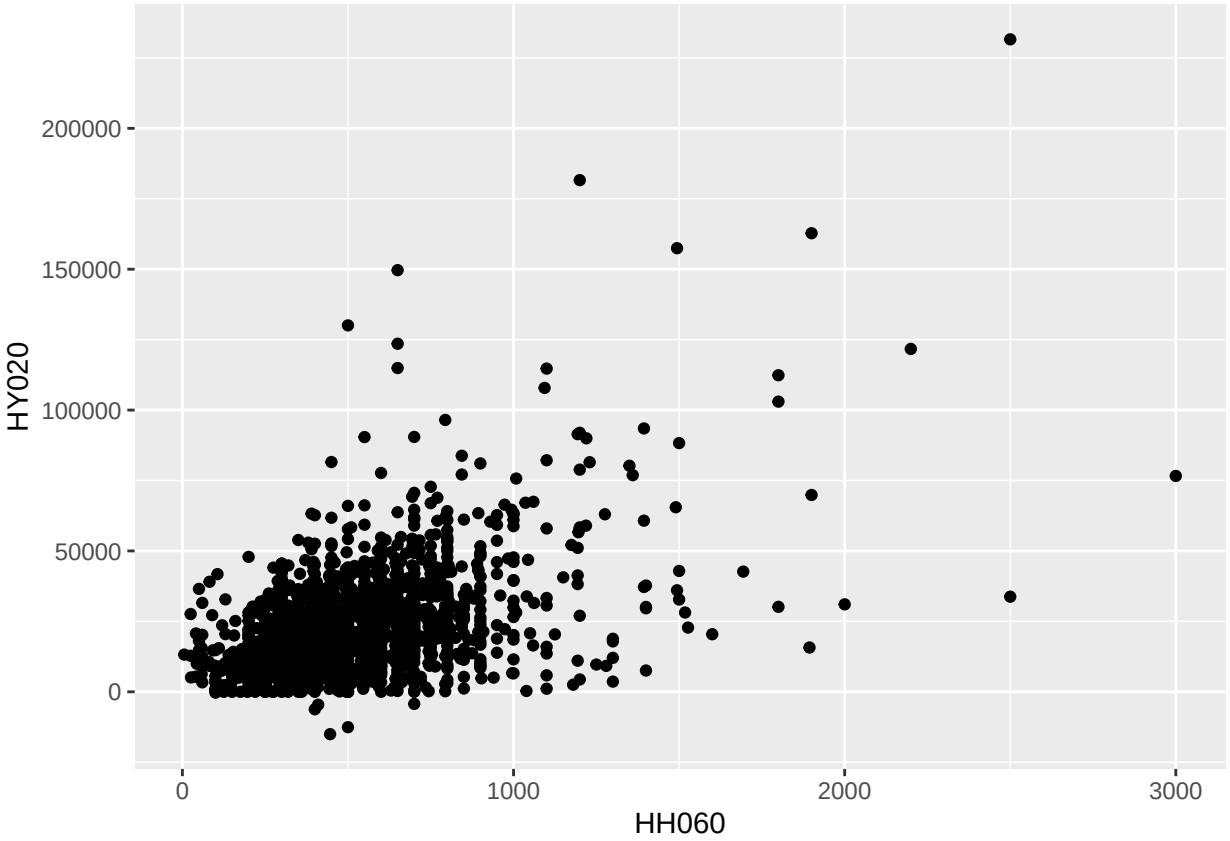
```
ggplot(data = subset(hogares.alquiler, !is.na(HH060)), aes(x = HH021, y = HH060, fill = HH021)) + geom_
```



Se observa que La mediana del valor a pagar por el alquiler de una vivienda corresponde a 456.08. Se puede observar que existen valores que superan el valor del tercer cuartil, por lo que se consideran como outliers, no se realizara la imputación de dichos valores ya que se entiende que existen ciertos casos en donde dependiendo de las características de las viviendas el valor del alquiler es muy elevado.

Precio del alquiler vs rentas de la vivienda

```
g1.hh060hy020 <- ggplot(data = subset(hogares.alquiler, !is.na(HH060))) +
  geom_point(aes(HH060, HY020))
g1.hh060hy020
```



Precio del alquiler vs Nacionalidad

Previo al análisis del precio del alquiler de viviendas por nacionalidad de los arrendatarios, se realiza la unión de los conjuntos datos de hogares y personas, con el fin de lograr identificar a las personas que residen en una vivienda alquilada.

```
# Se genera la variable con el identificador de la vivienda
personas <- personas %>% mutate(HB030 = as.integer(substr(PB030,1,nchar(PB030)-2)))

# Se genera un nuevo dataset resultado del join entre personas y hogares y se filtran solo las personas
personas <- left_join(personas, hogares.alquiler, by ='HB030', copy =FALSE)
personas.alquiler <- filter(personas, HH021 == 'En alquiler o realquiler a precio de mercado')

# Se genera el factor para verificar los labels d enacionalidad de las personas.
personas.alquiler$PB210 <- factor(personas.alquiler$PB210,levels =c(1,2,3), labels =c('España','Extranj
```

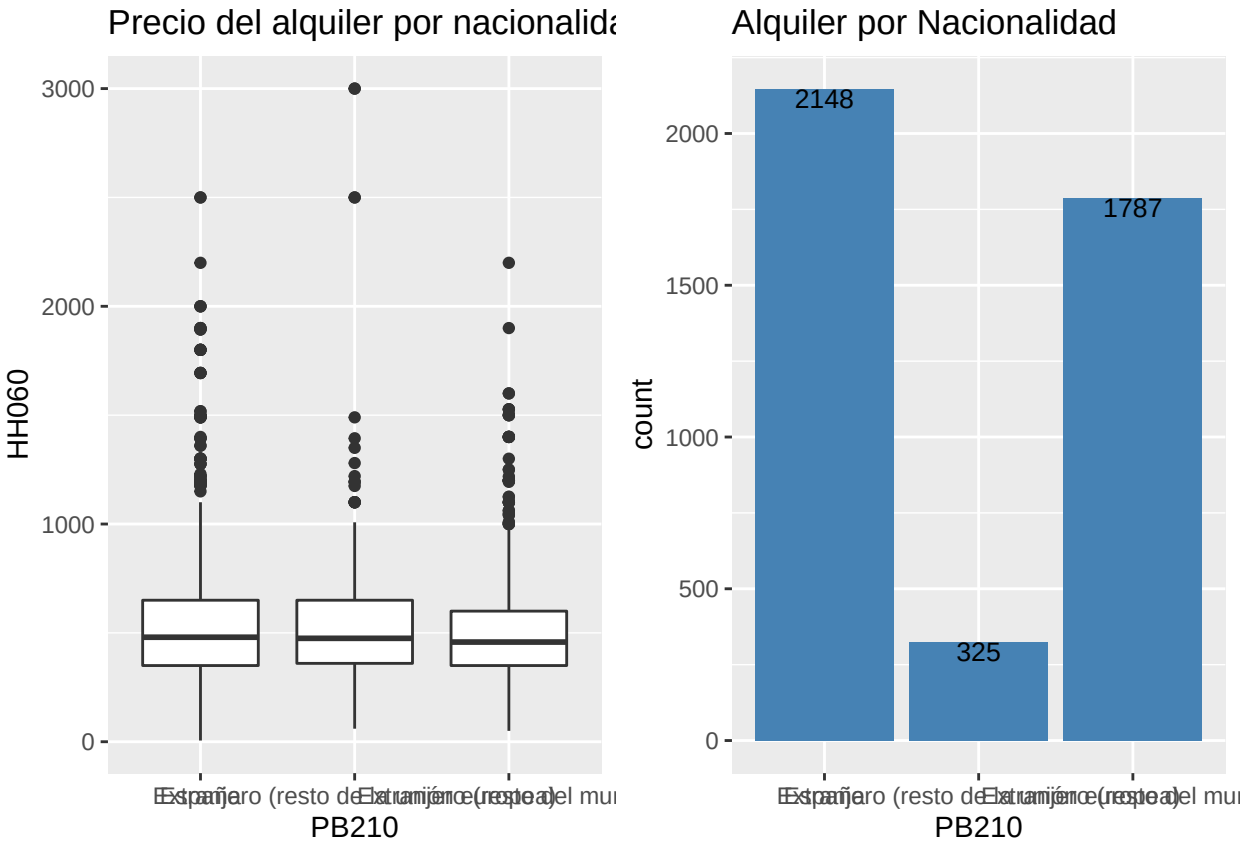
En la ECV, las nacionalidad se agrupa en 3 categorías que son: personas que nacidas en España, personas que nacieron en países miembros de la Unión Europea y personas nacidas en el resto de países dle mundo. Se analiza la distribución del precio de las viviendas por nacionalidad así como el número de personas que viven bajo un régimen de tenencial de alquiler por nacionalidad.

```
g.nacionalidad <- ggplot(data = subset(personas.alquiler, !is.na(PB210)), aes(PB210, HH060)) +
  geom_boxplot() +
  ggtitle('Precio del alquiler por nacionalidad')
```

```
g.nacionalidadCount <- ggplot(data = subset(personas.alquiler, !is.na(PB210)),aes(PB210) ) +
  geom_bar(fill="steelblue") +
  geom_text(aes(label=..count..), stat='count', vjust=1, color="black", size=3.5) +
  ggtitle('Alquiler por Nacionalidad')

grid.arrange(g.nacionalidad, g.nacionalidadCount, ncol=2)
```

```
## Warning: Removed 53 rows containing non-finite values (stat_boxplot).
```



Se puede observar que para los 3 grupos de tipos de personas residentes se tiene una medianda similar, lo que indica que la las personas que viven pagando un alquiler pagan casí lo mismo independientemente de su status migratorio. En el caso del número de personas, se tiene que las personas españolas cuentan con mas presencia en el mercado del alquiler con un total del 50,4%, en cambio las personas extranjeras que no pertenencen a países miembros de la Unión Europea cuentan con una proporción del 41,9% y por último, las personas extranjeras que vienen de países de la Unión Europea representan el 7,62%.

Precio del alquiler por el grupo de edad

En la ECV, encontramos el año de nacimiento de las personas encuestadas, a partir de dicho valor se calcula la edad a la fecha de la encuesta y se genera una nueva variable que contiene la edad de las personas agrupadas por las categorías: Menor de 30 años o mayor o igual a 30 años.

```
personas.alquiler <- personas.alquiler %>% mutate(edad = PB110 - PB140)
personas.alquiler <- personas.alquiler %>% mutate(grupos.edad = case_when(edad <30 ~ 'Menor de 30',
                                                                          edad >=30 ~ 'Mayor o igual a
```

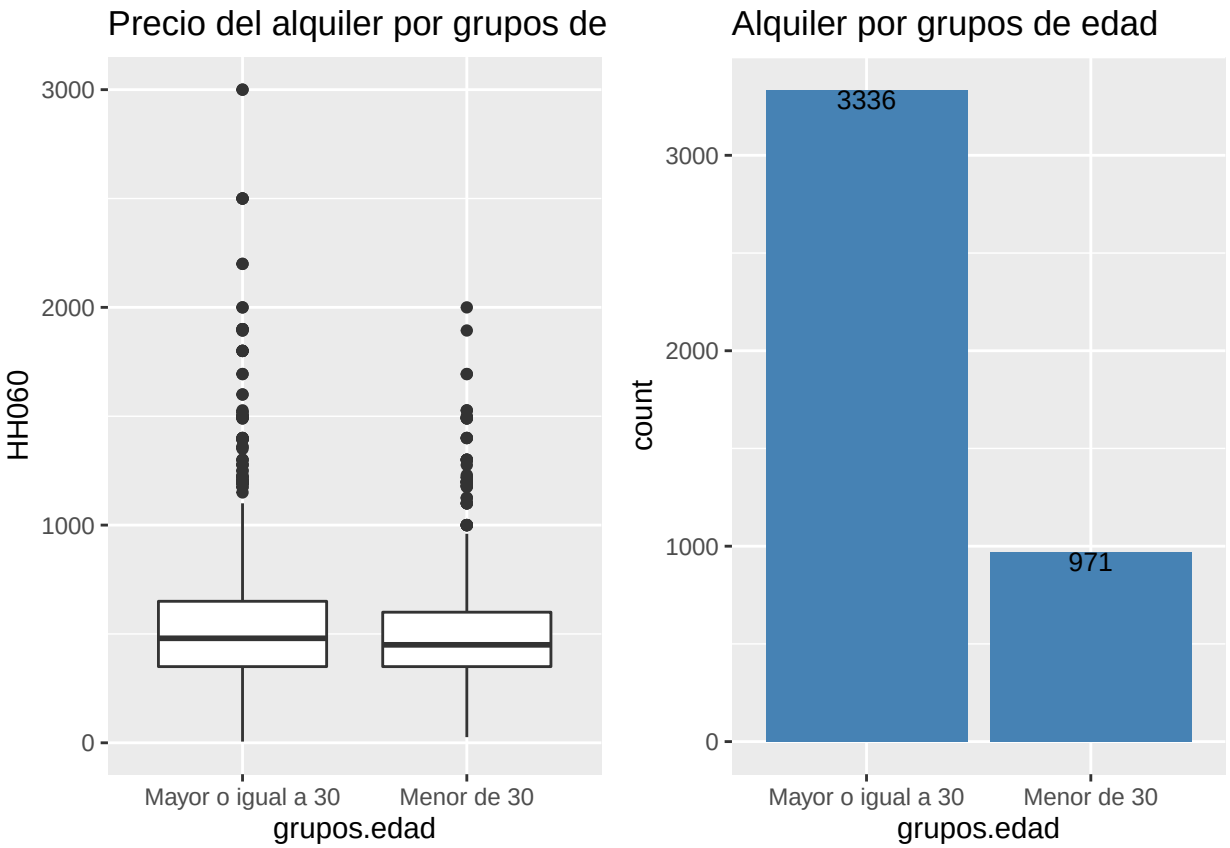
Se analiza mediante gráficos, cuál es la distribución del precio del alquiler de viviendas para las personas según su grupo de edad. Además, se analiza el número de personas por cada uno de estos grupos.

```
g1.Edad <- ggplot(data = subset(personas.alquiler, !is.na(grupos.edad)), aes(grupos.edad, HH060)) +
  geom_boxplot() +
  ggtitle('Precio del alquiler por grupos de edad')

g.edadCount <- ggplot(personas.alquiler,aes(grupos.edad) ) +
  geom_bar(fill="steelblue") +
  geom_text(aes(label=..count..), stat='count', vjust=1, color="black", size=3.5) +
  ggtitle('Alquiler por grupos de edad')

grid.arrange(g1.Edad, g.edadCount, ncol=2)
```

```
## Warning: Removed 57 rows containing non-finite values (stat_boxplot).
```



Se observa que independientemente de los grupos de edad, el precio del alquiler es similar, aunque en los hogares encuestados se observa una mayor presencia de personas con una edad igual o mayor de 30 años.

Número de habitantes por hogar

Verificamos a continuación la media del número de personas que habitan en un hogar con régimen de tenencia alquiler.

Alquiler por nivel de instrucción de las personas

```
# Se genera el factor para verificar los labels de los niveles de instrucción de las personas
personas.alquiler$PB210 <- factor(personas.alquiler$PB210,levels =c(1,2,3), labels =c('España','Extranj
```