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
```

En el presente proyecto se desea realizan un análisis sobre las fuentes de datos disponibles tanto oficiales como no oficiales con el fin de resolver las preguntas más importantes sobre el mercado del alquiler en España. Y con esto encontrar las principales patrones y también sesgos en este mercado.

Para ello se analizará en primer lugar la Encuesta de Condiciones de vida generada por el INE de manera anual. Ya que esta encuesta contiene información relevante de los principales aspectos de las condiciones de vida de los españoles, incluyendo el alquiler. Se eligió esta encuesta porque contiene información de las viviendas y personas que arriendan un inmueble para residir en el mismo así como de las viviendas y personas que ponen a disposición un inmueble para su alquiler.

En una segunda parte, se analizá la información disponible del portal inmobiliario Idealista, para así tener una perspestiva diferente de la que provee la encuesta y de esta manera poder comparar los resultados de las mismas.

Por último, se creará un modelo que permmita predecir el valor del alquiler de un inmueble a partir de los datos recogidos y del análisis realizado.

Encuesta de Condiciones de Vida - ECV

El objetivo general de la Encuesta de Condiciones de Vida (ECV) es la producción sistemática de estadísticas comunitarias sobre la renta y las condiciones de vida, que incluyan 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. Los microdatos que se distribuyen corresponden a la información transversal de la encuesta realizada en 2019.

La ECV está orientada a proporcionar informaciones comparables y armonizadas sobre los siguientes aspectos del nivel y condiciones de vida y de la cohesión social.

- Ingresos de los hogares privados. Situación económica.
- Pobreza, privación, protección mínima e igualdad de trato.
- Empleo y actividad
- Jubilaciones, pensiones y situación socioeconómica de las personas de edad.
- Vivienda, costes asociados.
- Desarrollo regional, movimientos migratorios.
- Nivel de formación, salud y efectos de ambos sobre la condición socioeconómica.

De los aspectos nombrados anteriormente, el referente a la vivienda y costes asociados son el objeto de estudio del presente proyecto, en especial para aquellas viviendas que constan con un régimen de tenencia de alquiler.

Extracción de datos

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
```

Limpieza de datos

Conjunto de datos de hogares Como primer paso en el proceso de data cleaning sobre el conjunto de datos de hogares, se realizará la eliminación de un conjunto de variables que son intrascendentes para el análisis que se desea realizar, ya que son datos genéricos corerspondientes a la encuesta como tal y no aportan con información significativa para el análisis. Estas variables se detallan en la tabla 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

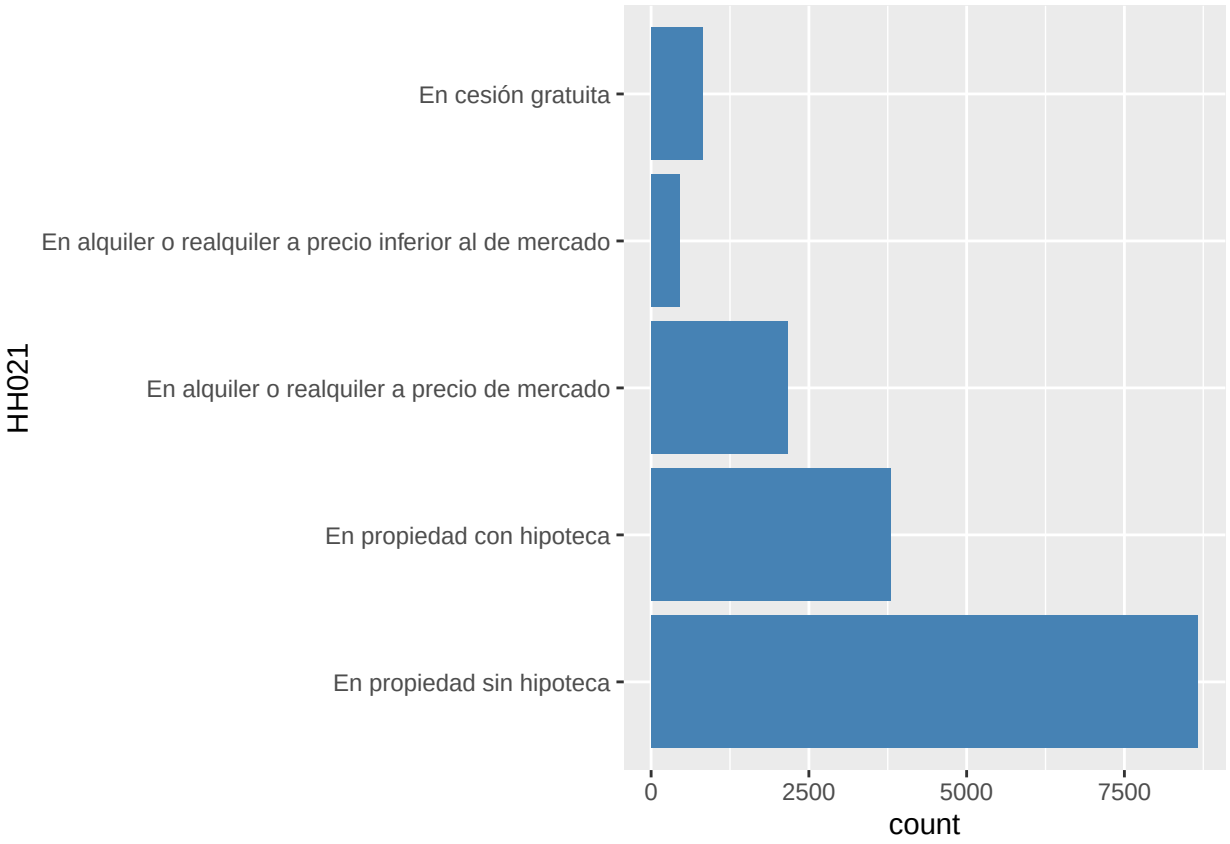
```
# Se genera un vector con las variables se eliminarán del data frame
col_borrar_hogares <- c('HB010','HB020','HB050','HB050_F','HB060','HB060_F','HB070','HB070_F','HB080','H

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

El conjunto de datos de la ECV correspondiente a los hogares contiene la variable `HH021`, la cual indica el régimen de tenencia de la vivienda en la que habitan las personas residentes en los hogares encuestados. Entre las categorías de dicha variable se cuenta la siguiente: *En alquiler o realquiler al precio del mercado.* Se filtrarán los datos del conjunto de datos de hogares según esta categoría ya que es la que interesa para el desarrollo del presente estudio. Presentamos el conteo de la variable `HH021` según sus categorias.

```
hogares$HH021 <- factor(hogares$HH021,levels = c(1,2,3,4,5),labels = c('En propiedad sin hipoteca','En p

g1 <- ggplot(hogares, aes(HH021)) +
  geom_bar(fill="steelblue") +
  coord_flip()
g1
```



Para la categoría de *En alquiler o realquiler a precio de mercado* se tiene un total de 2163 registros. El alquiler a precio de mercado corresponde a la cantidad que se debe pagar como contrapartida del derecho a utilizar una vivienda sin amueblar en el mercado privado, sin incluir los gastos de calefacción, electricidad, agua, etc.

Se filtra el conjunto de datos por la categoria antes mencionada.

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

Luego de filtrados los datos, se cuenta con un conjunto de datos de 2163 registros.

Una de las variables más importantes para el análisis del mercado del alquiler es aquella que contiene el valor pagado por el alquiler de una vivienda. Por ello es imprescindible revisar si estas variable contiene valores nulos y verificar la manera de imputar estos datos en el caso de ser necesario.

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

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

Tomando en cuenta que existen valores nulos, y con el fin de contar con datos consolidados se realizará la imputación de los datos vacíos en la variable HH060 que se refiere al valor del alquiler declarado. Se utilizará la imputación mediante regresión.

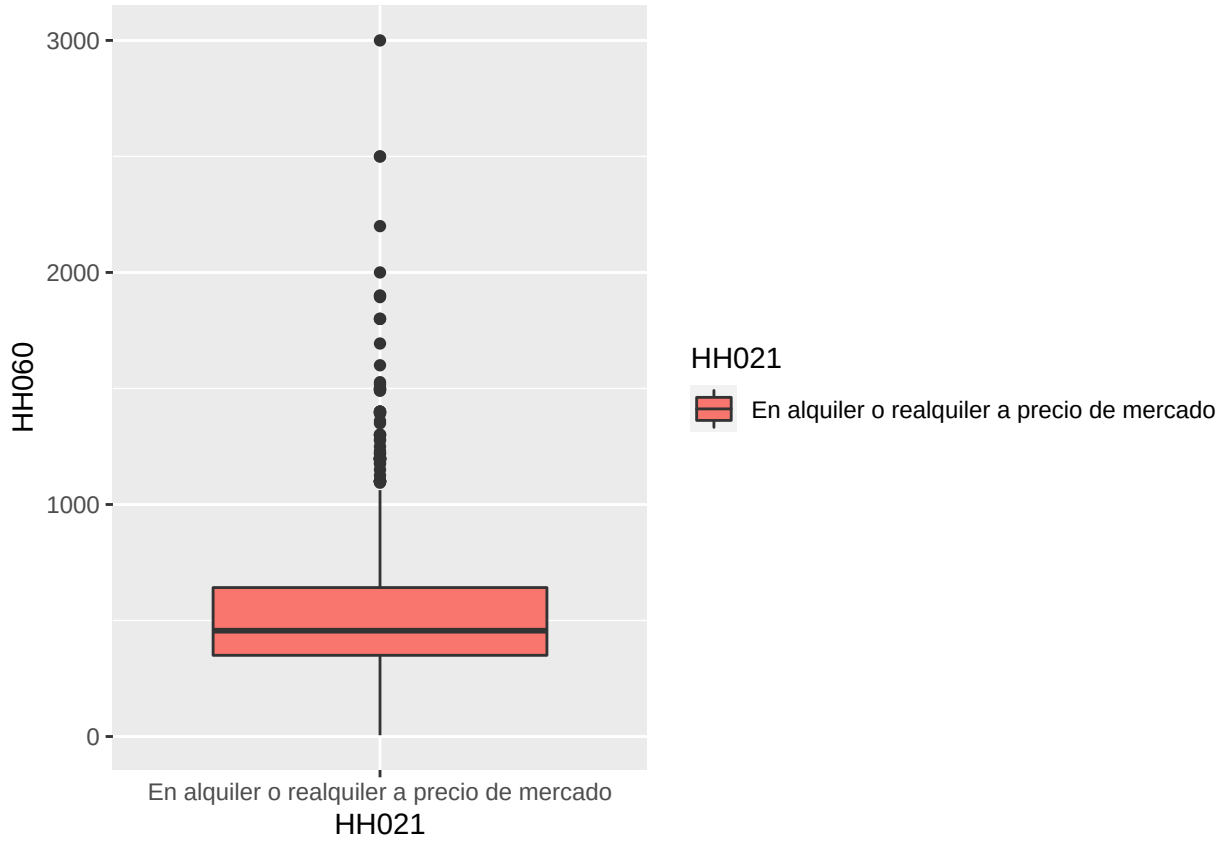
```
# Imputación aquí
```

Análisis de la información

Se realiza el análisis de la variable HH060, referente al valor pagado por el alquiler de una vivienda. Utilizando un diagrama de cajas se procede a revisar la distribución de la variable para el conjunto de datos.

```
ggplot(hogares.alquilados, aes(x = HH021, y = HH060, fill = HH021)) + geom_boxplot()
```

```
## Warning: Removed 32 rows containing non-finite values (stat_boxplot).
```




```
quantile(hogares.alquilados$HH060[hogares.alquilados$HH021 == 'En alquiler o realquiler a precio de mer
```

```
##      0%     25%     50%     75%    100%
##    5.50  350.00  456.08  641.89 3000.00
```

La mediana del valor a pagar por el alquiler de una vivienda corresponde a 456.08.

```r
hist(hogares.alquilados$HH060)
```