# 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(FSelector)
```

## 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= ',',
```

## 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:

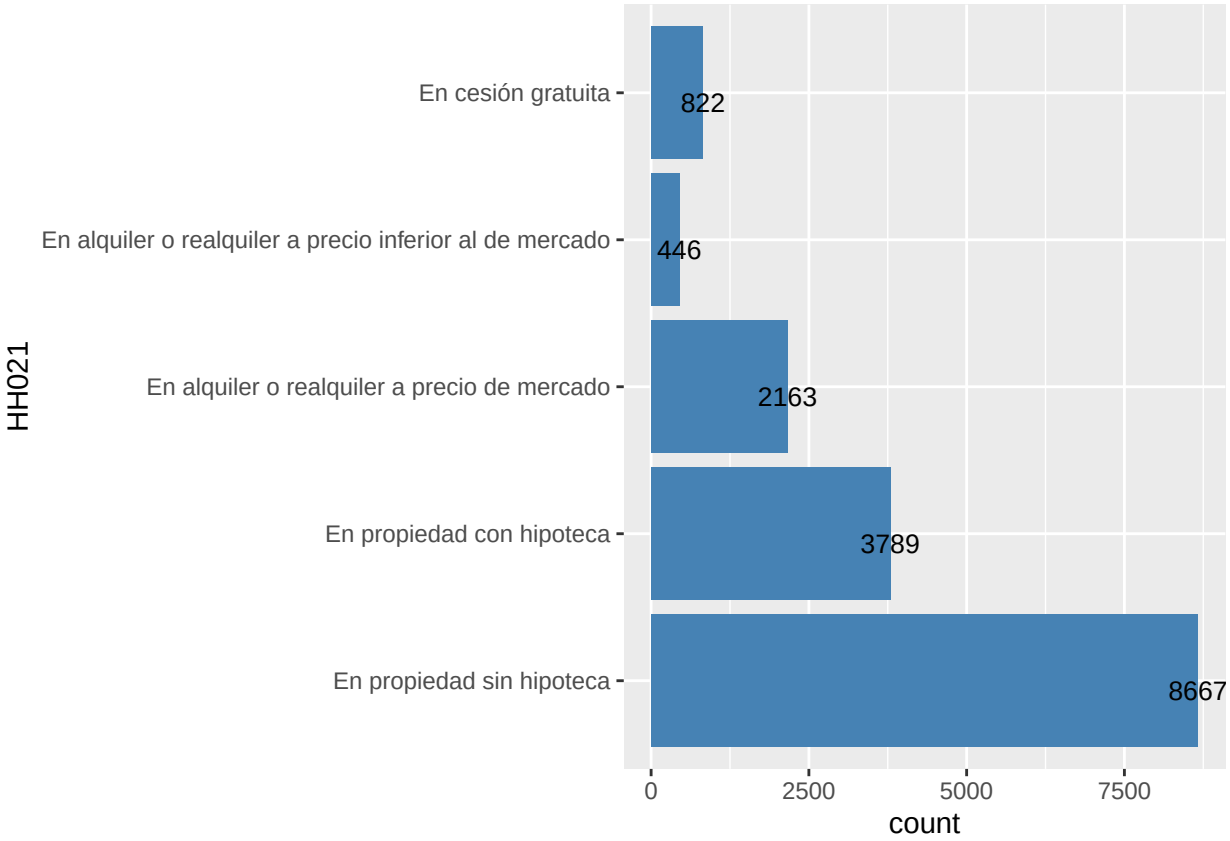
| Variable | Descripción |
|---|---|
| HB010 | Año de la encuesta |
| HB020 | País |
| HB050 | Mes de la entrevista |
| HB050_F | Mes de la entrevista - Flag |
| HB060 | Año de la entrevista |
| HB060_F | Año de la entrevista - Flag |
| HB070 | Identificación personal del informante del cuestionario 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

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

# 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, Para la tenencia de alquiler a precio de mercado se cuenta con un total de `2.163` registros y para el régimen de tenencia de alquiler a precio inferior al mercado se tienen 446 registros. Estos registros hacen referencia a hogares encuestados.