Proyecto Análisis Exploratorio de Datos

Adrián Lara Velasco, Andrea Romero Medina y Pablo Vicente Martínez

2023

Contents

Abstract

En este estudio de análisis exploratorio de datos, se examina un extenso conjunto de datos relacionados con hogares y personas, abordando diversas variables cruciales para comprender las dinámicas socioeconómicas. El análisis se centra en aspectos fundamentales de los hogares, como su tamaño, ubicación geográfica y modalidades de adquisición de vivienda. Además, se profundiza en la caracterización de las personas que conforman estos hogares, explorando variables como la edad, nacionalidad y nivel educativo. A través de técnicas estadísticas descriptivas y visualizaciones, se revelan patrones, correlaciones y tendencias emergentes en los datos, arrojando luz sobre las complejas interrelaciones entre los factores estudiados. Este enfoque proporciona una visión integral que no solo destaca la diversidad de contextos residenciales, sino también la heterogeneidad de la población, ofreciendo valiosas perspectivas para la formulación de políticas y la toma de decisiones informada en temas relacionados con la vivienda y el bienestar social.

Carga de los datos

Los conjuntos de datos a estudiar se obtuvieron de dos archivos ".csv" en la siguiente url del INE

(*https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176952&menu=resultados&secc=1254736195203&idp=1254735572981#!tabs-1254736195203*)

La carga de los mismos fue bastante sencilla ya que el porpio INE proprocionaba **tiny datasets**.

Los conjuntos de datos con los que inicialmente trabajamos se denominan hogar y persona (hogar_2 y persona_2 se explicarán más adelante). hogar es un conjunto de 88783 observaciones y 27 variables mientras que persona presenta 220198 observaciones y 27 variables.

Hay una serie de variables presentes en ambos conjuntos que actuan como primary keys para su posterior unión (codebook).

Limpieza de los datos

Antes de empezar a trabajar en los dataset, miramos la estructura de las variables para obtener una primera referencia sobre los datos.

```
str(hogar)
str(persona)
```

Podemos observar que tenemos muchas variables tipo *integer* así como algunas variables de tipo *numérico*, *character* y algunas de tipo *lógico*. En este vistazo inicial también se observa una gran presencia de valores ausentes (NA's) en determinadas variables.

El INE nos ofrece un informe con la metodología empleada para la obtención de los datos así como la explicación de las variables que se encuentran en el dataset. Para recoger la información se utilizaron cuestionarios donde el participante debe contestar a cada pregunta con una respuesta dentro de una lista de respuestas predefinidas. Además, hay preguntas que no todos los participantes deben contestar porque dependen de respuestas anteriores, por tanto esto nos aporta información en dos direcciones:

1º. Gran parte de las variables de los conjuntos de datos son categóricas

2º. Muchos datos ausentes se deban a no tener que responder una determinada pregunta

Al ser una gran cantidad de variables de tipo *categórico*, estas podían aparecer en forma de texto (factor), o en forma de número (integer). Ambos enfoques tienen sus ventajas e inconvenientes. Si tenemos las variables en formato texto, como factor, podemos saber de manera más sencilla qué significa cada dato, lo que permite clasificar los datos con mayor facilidad, encontrar relaciones etc. También nos permite añadir categorías en formato texto alla donde hay NA's, como categorías adicionales a las existentes, y poder operar con un dataset sin valores ausentes. La principal desventaja es que no se puede modelizar con variables tipo texto, necesitan ser transformadas a tipo numérico para ello. La ventaja de utilizar variables de tipo numérico es la que se acaba de presentar, el poder modelizar. La principal desventaja es el hecho de que en algunas ocasiones, como se explicará posteriormente, no se pueden imputar valores a los NA simplemente porque no procede hacerlo.

Por tanto se ha decidido utilizar dos datasets, uno donde transformamos las variables categóricas a texto, y otro dataset donde mantenemos las variables en un formato numérico, exceptuando algunas donde sería muy complicado de operar y es preferible operar con texto, por ejemplo la variable 'Nombre de países'.

- hogar y persona son datasets donde transformamos las variables a tipo texto
- hogar_2 y persona_2 datasets donde mantenemos el formato numérico.

Para este proyecto trabajaremos principalmente sobre el dataset con variables de tipo texto pero se crea el de formato numérico como conjunto de datos que se podría usar para modelizar (encoding).

Por tanto vamos a transformar las variables deseadas a formato texto.

Transformación de variables

Para transformar las variables utlizamos los diccionarios adjuntos al dataset, donde se explica cómo se llama cada variable, qué significa y cómo está codificada.

Cada hoja del diccionario contiene información sobre las variables. Esta información es obtenida a partir de la ejecución de una fucnión y posteriormnete sustituida en el dataset de trabajo.

Repetimos este proceso para el conjunto de datos de persona/persona_2

De este modo hemos trasfromado las variables que queriamos tener en texto y mantenemos aquellas variables númericas de interés para poder tener un df utilizable para modelizar y otro para analizar cualitativamente.

```
head(persona)
head(hogar)
head(persona_2)
head(hogar_2)
```

Unimos finalmente los subconjuntos de datos para tener el conjunto que vamos a utilizar en el análisis.

- df: Donde almacenamos la mayor parte de las variables en tipo texto
- df_num: Donde almacenamos la mayor parte de las variables en tipo numérico

Obervamos la estructura del df y observamos una variable (PERIODO) que presenta un formato de tipo integer, pero que realmente representa el año y el trimestre de toma de los datos, por lo que se ajusta más a un tipo fecha. Por lo tanto modificamos este dato para expresarlo en dicho tipo.

Las variables de tipo *character* las podemos trasnformar a factor, sin embargo nos vamos a esperar a haber realizado primero el análisis de NA's ya que vamos a crear nuevas categorías de datos al imputar valores.

Análisis de NA's

El dataset contiene numerosos NA's, esto se puede deber a diversos factores como valores perdidos, o que no proceda la existencia de ningún valor en ese caso. Este dataset contiene mayoritariamente NA's debido a lo comentado previamente sobre que hay algunas preguntas a las que n0o corresponde contetsra dependiendo de las respuestas dadas previamente. Por tanto, si analizamos el por qué de cada caso, podremos obtener información adicional.

Como veremos a continuación, no tiene sentido asignar un valor numérico a un valor NA, por tanto para nuestro dataset numérico no tiene sentido imputar valores a los NA. Si se quisiera modelizar o utilizar cualquier técnica que requiera cálculos, es imprescindible tener el dataset adaptado a cada modelo que se quiera implementar, aplicando distintos filtros que hace que desaparezcan los NA. Por ejemplo, filtrar la edad en valores mayores de 16 años. Esto se analizará con más detalle a lo largo de la sección.

Hay variables con un 100% de NA y otras con porcentajes del 50%, 60%, 45%, 16%, etc. Un aspecto interesante que podemos observar, es que hay variables que tienen exactamente el mismo porcentaje de NA's, esto podría indicarnos relaciones entre las variables.

Las primeras variables que vamos a analizar son P01:P19. Estas indican las relaciones de parentesco entre los miembros del hogar.

```
vector_p <- paste0("P", sprintf("%02d", seq(1, 19)))
NA_analisis_P <- sapply(df[,vector_p], function(x) {mean(is.na(x)) * 100})
NA_analisis_P
```

```
##       P01       P02       P03       P04       P05       P06       P07       P08
##  40.31962  70.73043  87.98036  96.97454  99.13623  99.72116  99.89737  99.95777
##       P09       P10       P11       P12       P13       P14       P15       P16
##  99.98002  99.99046  99.99637  99.99955 100.00000 100.00000 100.00000 100.00000
##       P17       P18       P19
## 100.00000 100.00000 100.00000
```

Vemos que casi todas ellas tienen porcentajes muy elevados de NA's, inclusive del 100%. Esto es lógico pues es poco común que hayan familias de más de 4 miembros, además solamente con P01 ya podemos tener la información sobre las relaciones de parentesco de todos los miembros de la familia por lo que podemos eliminar P02:P19 sin perder información.

Si le imputamos al miembro que realizó la encuesta, NPV == 1, el valor "Miembro encuestado" (lo cual tendría sentido), y eliminamos el resto de variables, vemos que realmente no tenemos valores ausentes. Esta tendencia de, mediante una sola operación de imputación razonable, arreglar los valores ausentes, es recurrente en el dataset.

A continuación se estudian todas aquellas variables que tienen relación con el nucleo familiar

(EC,NHIJOME_NUCLEO,NHIJO_NUCLEO,SITUNUCLEOFAM,PAREJA,NUCLEOFAM, NHIJOMENOR,ECPAR,NACPAR,SEXOPAR,HIJOSDEAMBOS,NHIJOPAR,NHIJO_NUCLEO,NHIJO)

Todas ellas estan relacionadas ya que la información que contienen varía ligeramente, sin embargo, observamos que si los individuos que contestaron el cuestionario en el apartado de PAREJA establecieron que "No convive en pareja", hacía que se quedara sin contestar, pues no se puede contestar algo sobre el EC de la pareja sin tenerla.

Para las variables **NUCLEOFAM**, **NHIJOPAR** y **NHIJO**, se han tenido que realizar imputaciones de datos, pues en principio no hay ningún criterio conocido por el que esos datos no estuvieran presentes. Se buscaron variables que contengan información muy similar o igual a la información que recoge la variable. Al tener el dataset una gran cantidad de variables con cambios marginales entre una y otra, podemos imputar datos de una variable a otra de forma relativamente razonable, además el número de datos perdidos en este caso es relativamente pequeño (no más de un 3% de NA reales de esas variables).

La lista completa de relaciones se puede visualizar en el código.

Seguidamente podemos ver que las variables **ESTUDIOS** y **RELACT** también tienen valores ausentes relacionados. Estas preguntas del cuestionario solo se tomaron a los mayores de 16 años, por tanto si filtramos el df por edad, vemos que los NA's desaparecen.

De forma análoga, aquellos que contestaron que no se encontraban "Trabajando a tiempo completo" o "Trabajando a tiempo parcial", no correspondía que contestaran nada en el apartado de **OCUPA**, ya que no trabajan.

Respecto a las variables relacionadas con el **lugar de nacimiento** y la **nacionalidad**, hay apartados en los que correspondía responder en función o no de la nacionalidad del encuestado, por ejemplo, los ciudadanos nacidos en España no debían contestar cuándo llegaron a España o cuándo obtuvieron la nacionalidad. Qué valores ausentes había y cómo se han imputado se puede ver en el código abajo:

Finalmente quedaría la variable **ANEDI**, que solo registra el año de construcción del edificio a partir del año 2000. Por tanto, todo edificio construido previamente no se incluyó. Sin embargo, tenemos variables que tienen esta información y más, por lo que al ser una variable que apenas aporta información relevante, podemos eliminarla.

Una vez hecho todo el proceso de busqueda de datos perdidos e imputación, podemos ver como para nuestro conjunto de datos ya no tenemos datos ausentes, de hecho en ningún momento tuvimos (salvo en algún caso muy puntual). Sin embargo, era necesario añadir nuevas categorías dentro de las variables que capturasen esa información implícita.

El motivo por el que no se han podido imputar al df numérico todos esos NA's es porque ningún tipo de imputación numérica hubiera sido realmente cierta, y hubiera sesgado los verdaderos valores del conjunto. Un

ejemplo ilustrativo es la edad de llegada a España, para un ciudadano nacido en España podríamos imputar el valor 0, sin embargo, esto se interpretaría como: "Ha llegado hace 0 años", lo cual sesgaría la distribución de años hacía 0, pues la mayoría de personas que contestó la encuesta eran de nacionalidad española.

Este caso es el mismo para todas las variables con valores ausentes, por lo tanto, para no hacer un posterior análisis erroneo, es importante realizar un filtrado correcto de los valores ausentes antes de realizar cálculos numéricos. Primero, porque si no R no será capaz de realizar cálculos con datos ausentes, pero segundo, y más importante, porque al ser NA's donde no se pueden imputar datos, se nos esta indicando implícitamente que no debemos utilizar más datos de los que hay, ya que si no estaríamos contestando preguntas con información de sujetos que no tienen nada que ver con la misma.

```
NA_analisis <- sapply(df, function(x) {mean(is.na(x)) * 100})
NA_analisis
```

```
##         ID_VIV        PERIODO         TAMANO         IDQ_PV             CA
##              0              0              0              0              0
##         FACCAL          REGVI         COCINA          ASEOS      COMEDORES
##              0              0              0              0              0
##    DORMITORIOS      TRASTEROS       OTRASHAB          HABVI       METROSVI
##              0              0              0              0              0
##        TIPOVIV          FEDI        TAMTOHO     DENSIDADVI         TIPOHO
##              0              0              0              0              0
##          NACHO      NUCLEOFAM       NHIJOPAR          NHIJO     NHIJOMENOR
##              0              0              0              0              0
##            NPV           SEXO           EDAD             EC          NACIM
##              0              0              0              0              0
##        PNACIMT      EDADFLLEG       TPOFLLEG            NAC          PNACT
##              0              0              0              0              0
##    NACNACIMESP     PNACNACIMT     TPOFNACESP     NACIMPADRE   PNACIMPADRET
##              0              0              0              0              0
##     NACIMMADRE   PNACIMMADRET            P01       ESTUDIOS         RELACT
##              0              0              0              0              0
##          OCUPA         SITUHO       SITUHO_D         PAREJA        SEXOPAR
##              0              0              0              0              0
##         NACPAR          ECPAR   NHIJO_NUCLEO NHIJOME_NUCLEO   HIJOSDEAMBOS
##              0              0              0              0              0
##  SITUNUCLEOFAM
##              0
```

Cambios de tipo de variables

Como mencionamos anteriormente, vamos a cambiar las variables *character* a *factor*, no lo habiámos hecho antes por la imputación de valores que hemos hecho en el apartado anterior, de esta forma podemos ver cuantos levels tenemos para cada variable con un simple vistazo a la función str.

```
df <- df |>
  mutate_if(is.character, factor)

df_num <- df_num |>
  mutate_if(is.character, factor)

str(df)
```

Analizando los levels de las variables factor, nos encontramos que para las variables relacionadas con paises, aparecen levels con código "966" y "555", que no aparecen en los diccionarios proporcionados en el INE y que por tanto se han debido de tratar de errores de escritura, por tanto vamos a filtrarlos.

Cambio nombre variables

Cambiamos también de nombre algunas variables para mejorar la comprensión de las mismas a la hora de leer el set de datos. Algunos de estos cambios son IDQ_PV -> PROVINCIA o EDADFLLEG -> EDAD_LLEG_ESP.

Recodificación de variables

En un inicio, nuestro dataset presentaba un tipo de codificación ya establecido. El problema era la dificultad de interpretar el dataset y sus posibles relaciones, entonces se decidió descodificar primero las variables para poder entender qué significaba cada variable de forma sencilla. Con esto vimos que el análisis de hecho se podía realizar de forma más intuitiva y decidimos que nuestras variables podían permanecer en formato *character*, sin embargo nos vimos en la necesidad de mantener un dataset con principalmente variables numéricas con la que se pudiera eventualmente modelizar por ejemplo.

Una vez investigado, observamos que la codificación de algunas variables no era la adecuada por el tipo de variable categórica. Las variables categóricas se dividen en dos grupos: ordinales y cardinales. En las variables categóricas ordinales, se pueden codificar acorde a un orden númerico, ya que existe una jerarquia. a > b > c. En las variables cardinales, por el contrario, no se aprecia un orden entre las variables: Andalucia !> Aragón. Esto es importante porque para cada caso las técnicas de codificación son distintas. Las variables categóricas ordinales estan bien codificadas en el dataset, las categóricas ordinales por el contrario han requerido una recodificación, nosotros hemos utilizado "One-hot encoding", también conocido como convertir a "dummy variables" que basicamente es un tipo de codificación binaria. 1 es que la variable toma un valor, 0 que no lo toma.

Como variables ordinales en este caso podemos considerar a **TAMAÑO**, **ESTUDIOS**, **TIPOVIV**, **FEDI**, mientras que el resto son cardinales.

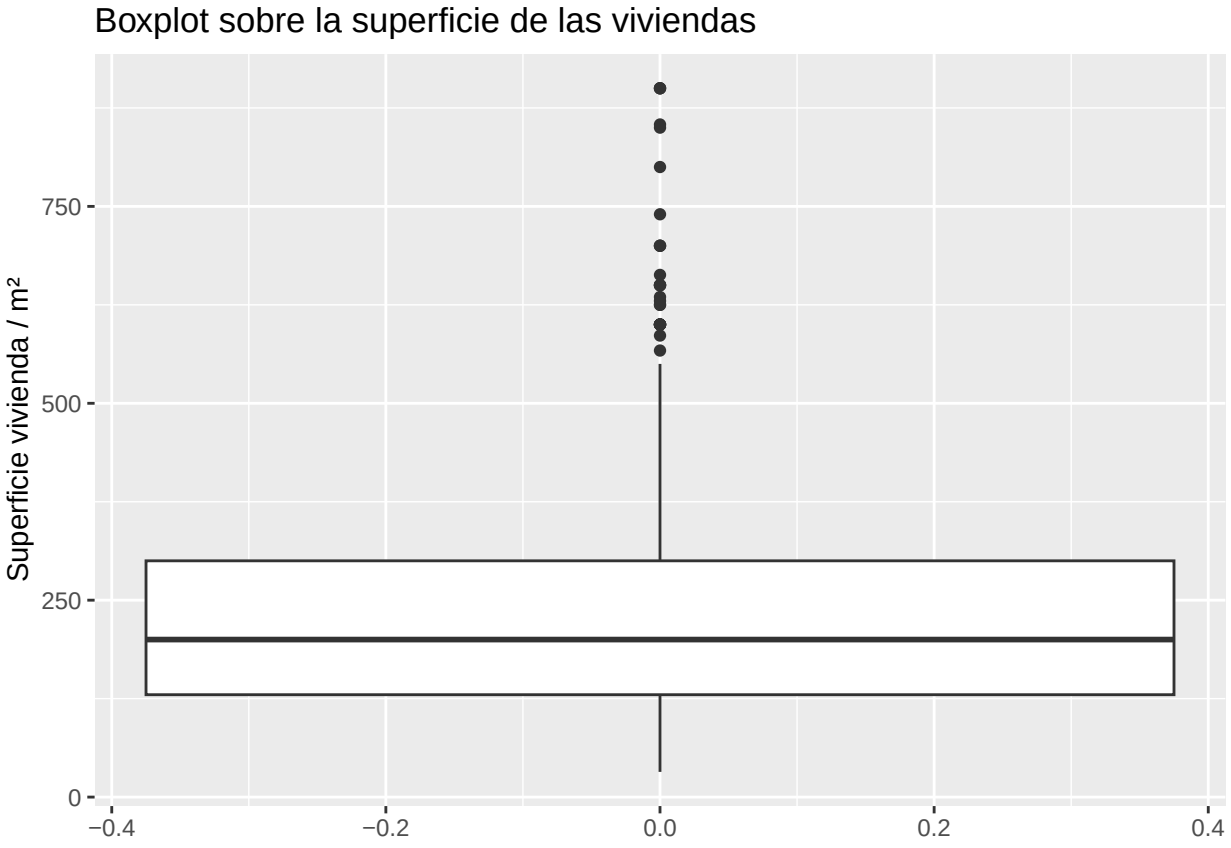
```
columnas_dummy <- c("REGVI","NAC_HO","SEXO","EC","NACIM","COCINA","NAC",
"NACNACIMESP","NACIMPADRE","NACIMMADRE","OCUPA","PAREJA","SEXOPAR","NACPAR",
"ECPAR","HIJOSDEAMBOS")

df_num_fctr <- df_num %>%
  select(all_of(columnas_dummy)) %>%
    mutate_all(factor)

dummy <- dummyVars(paste("~", paste(columnas_dummy,collapse = " + ")),
                   data = df_num_fctr)

df_dummies <- data.frame(predict(dummy, newdata = df_num_fctr))

df_merged <- cbind(df_num, df_dummies)

df_merged <- df_merged %>%
  select(!(columnas_dummy))

df <- df %>%
  mutate(METROSVI = as.numeric(METROSVI))
```

Detección de anomalías

Durante el proceso de detección de anomalías nos hemos servido de una serie de principios muy básicos. Al estar nuestro conjunto de datos repleto de datos categóricos, es complicado determinar que supone una anomalía y que no, en este caso nos hemos ceñido a las variables que mediante el sentido común se puede determinar que constituyen anomalías del conjunto, se han utilizado counts sobre las habitaciones del hogar para detectar posibles valores fuera de lo normal, ya que si aplicaramos este principio sobre otras variables como **NAC** o **EC** podriamos pensar que "No ser español" o "Viudo/a" podrían ser anomalías y eliminarlas del conjunto nos quitaría riqueza e información relevante del conjunto.

Detección de outliers

Apenas tenemos dos variables puramente numéricas que podamos utilizar de forma efectiva en la detección de outliers, **METROSVI** y **DENSIDADVI**, para la detección de outliers podemos utilizar técnicas visuales simples como un Boxplot o podemos utilizar formulas como 3sigma, Hampel, Boxplot entre otras opciones. En este caso hemos usado varios enfoques, primero visualizar nuestros conjuntos para ver si tenían o no outliers y luego aplicar los métodos de detección de outliers. El criterio que se aplica para determinar la elección del método es unicamente observar que datos nos indica y utilizar el sentido común de que nos parece o no outlier, sin embargo este criterio es subjetivo y cualquiera podría elegir un criterio diferente al utilizado aquí. En el caso de **METROSVI**, se ha decidido utilizar el método 3sigma dado que el resto de métodos resultan demasaidos restrictivos (outliers a partir de $180m^2$), y para **DENSIDADVI** se ha decidido utilizar el método de boxplot ya que otros métodos, como 3sigma, ofrecen límites demasiado laxos.

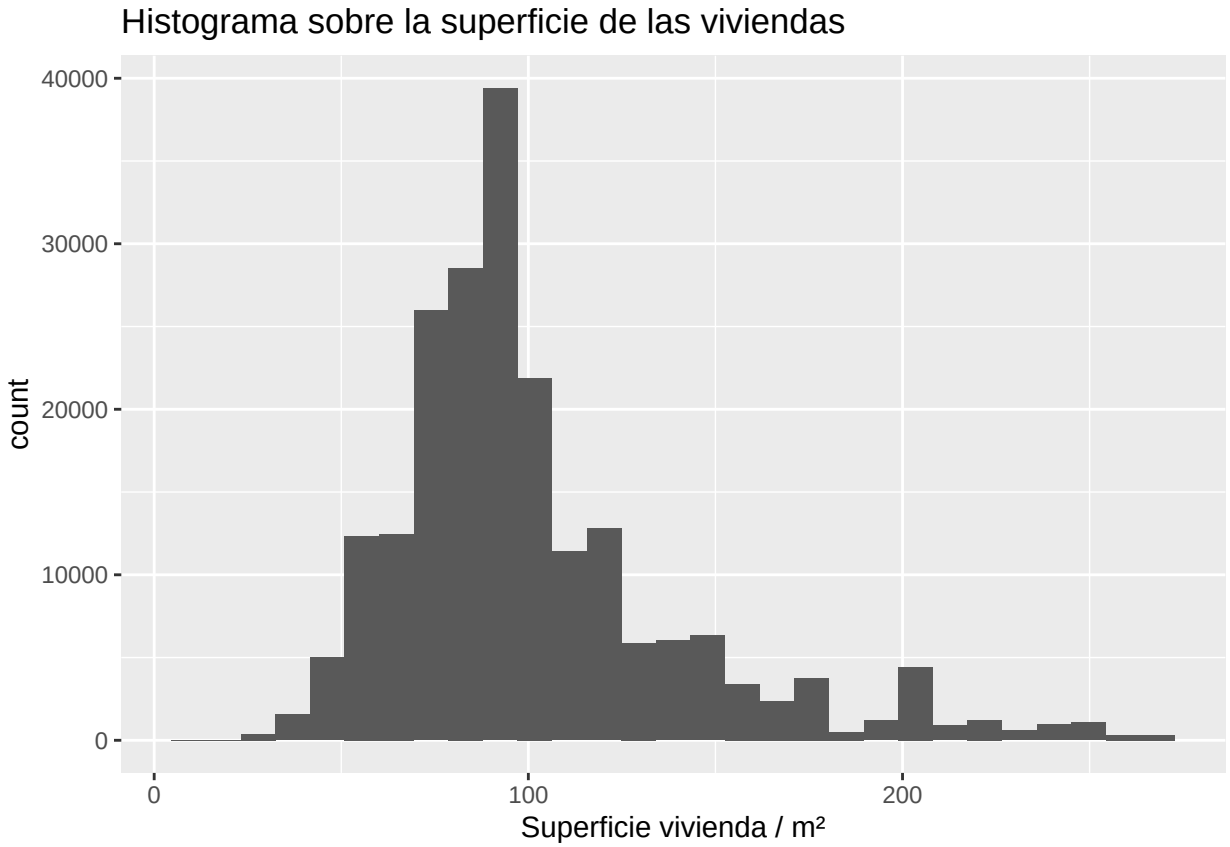


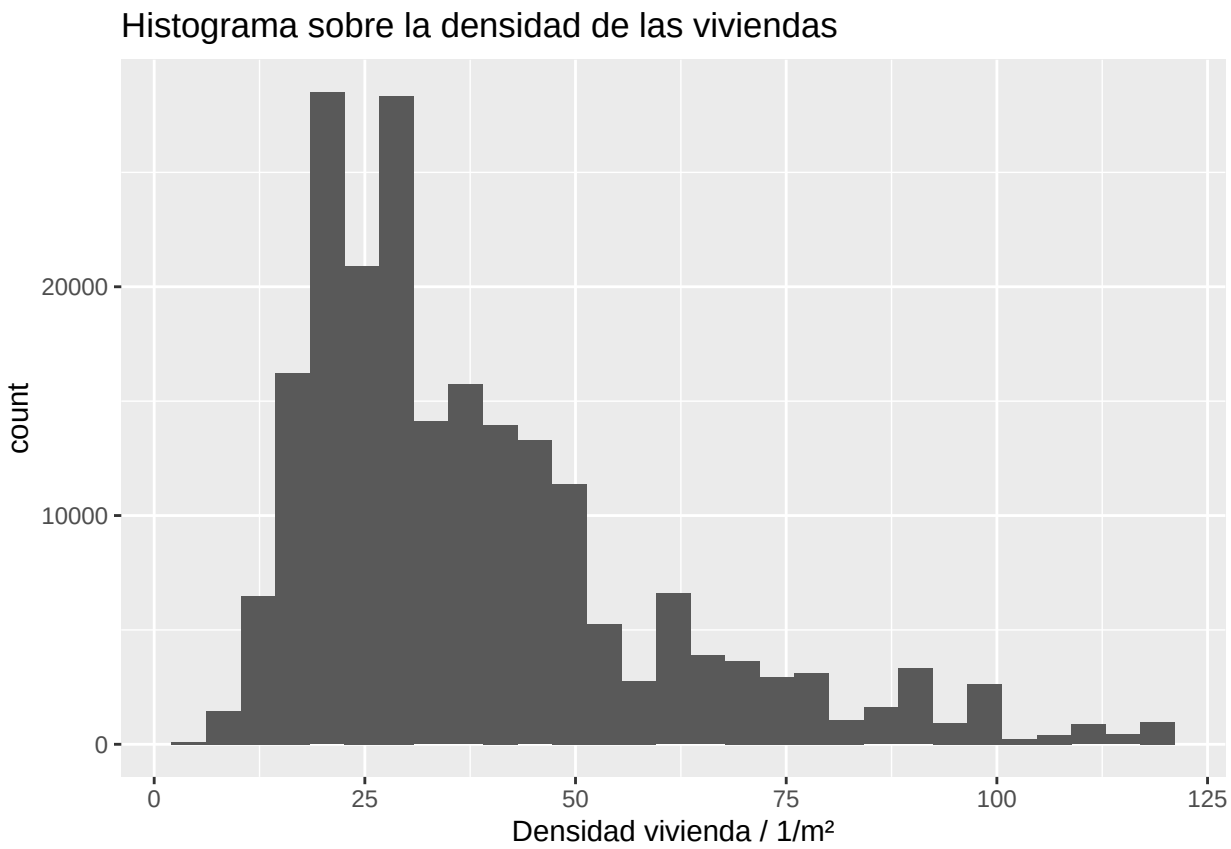

```
## El resultado para la detección de outliers en la variable de m² de las viviendas es:

##          method    n nMiss nOut    lowLim    upLim minNom maxNom
```

```
## 5%       p5-p95 1157     0  107   80.0000 460.0000     82    456
## 1     tresSigma 1157     0   19 -160.2066 620.2896     32    600
## 11       Hampel 1157     0   28 -164.7196 564.7196     32    550
## 12 ReglaBoxplot 1157     0   28 -125.0000 555.0000     32    550
```

El resultado para la detección de outliers en la variable de m^2 por persona es:

```
##          method    n nMiss nOut     lowLim    upLim minNom maxNom
## 5%       p5-p95 1157     0  116   24.96000 261.0000   25.0  260.0
## 1     tresSigma 1157     0   21 -150.00616 345.5693    7.9  327.0
## 11       Hampel 1157     0   91  -74.81174 221.4117    7.9  219.0
## 12 ReglaBoxplot 1157     0   83  -67.50000 232.5000    7.9  230.5
```


Histograma sobre la superficie de las viviendas
40000
30000
20000
10000
0
count
0
100
200
Superficie vivienda / m²



Hemos decidido mantener el dataframe que contiene outliers ya que podría ser interesante si alguien quisiera observar si existen diferencias en el resto de variables del conjunto entre el conjunto de datos filtrado de outliers y anomalías y el conjunto que solo tiene estos valores anormales. Para este proyecto sin embargo no vamos a analizar este punto.

Posibles preguntas a plantear

1. ¿El tamaño de la provincia afecta a las carácteristicas del hogar?
2. ¿Afecta tu nivel de estudios en donde vives?
3. ¿Tu estado civil tiene relación con tu hogar?
4. ¿Existen diferencias entre nacionales y extranjeros en nivel de estudios, ocupación, hijos, tamaño de población o tipo de vivienda?
5. ¿Está correlacionada la edad con tu vivienda?
6. ¿Existe alguna diferencia entre los extranjeros nacionalizados y los no nacionalizados?

Casos a estudiar

Análisis univariante

La mayoría de las variables con las que estamos trabajando son categóricas, variables numéricas solamente tenemos *DENSIDADVI* y *METROSVI*. Calcularemos algunos de los estadísticos más comunes para *METROSVI*, ya que nos podrán ayudar a una mejor comprensión de los datos:

```
## La media de metros cuadrados por vivienda es: 107.0207
```

```
## La mediana de metros cuadrados por vivienda es: 90
```

```
## La desviación típica de los metros cuadrados por vivienda es: 62.85405
```

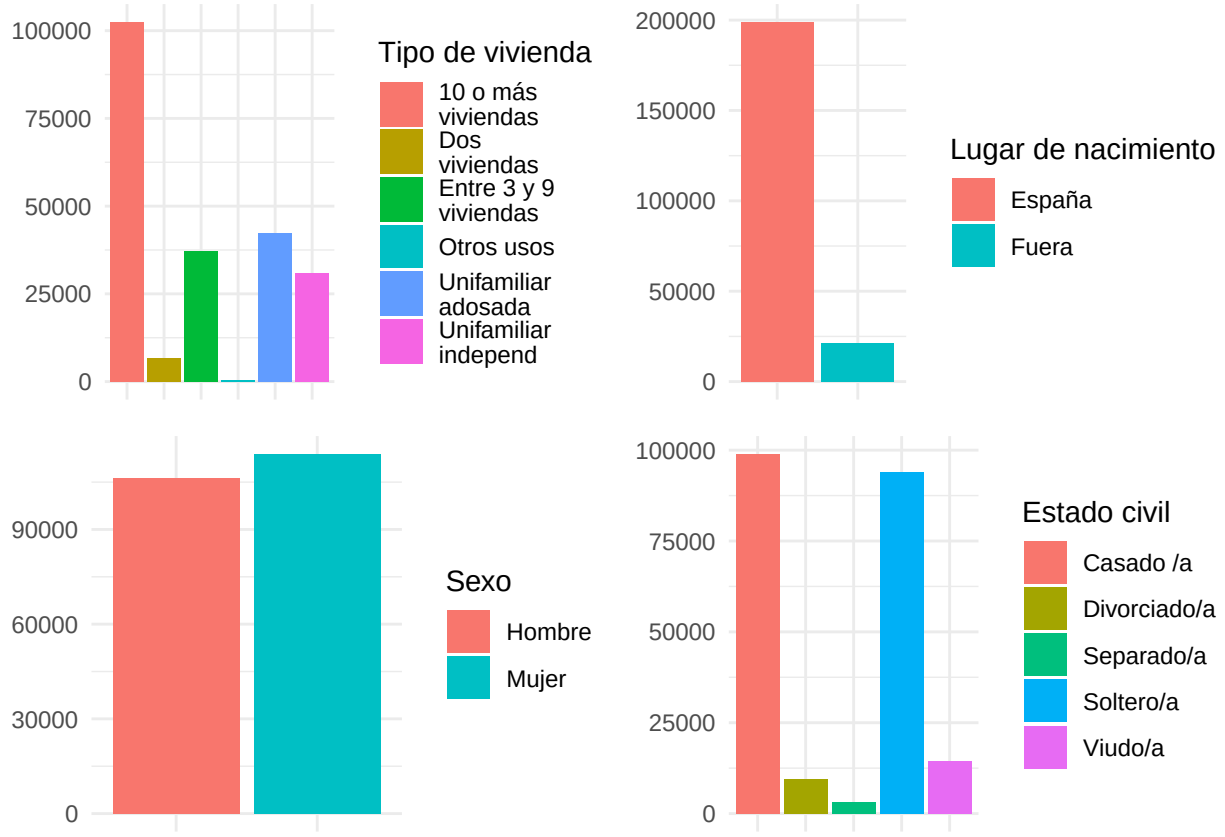
Que la mediana sea menor que la media quiere decir que hay más viviendas con menos metros cuadrados, y las casas con más metros cuadrados, que aumentan el valor de la media, su diferencia de metros debe ser significativa.

En este caso tenemos una desviación típica de 62.85, lo cual sugiere que los tamaños de las viviendas tienden a alejarse de la media en una cantidad considerable, lo que podría deberse a la presencia de viviendas más grandes y más pequeñas en el conjunto de datos.

En el análisis bivariante estudiaremos como las variables *DENSIDADVI* y *METROSVI* están correlacionadas. Esto se debe a que la variable *DENSIDAVI* la hemos obtenido del cociente entre *METROSVI* y *PERS_HOGAR*.

Para algunas de las variables categóricas lo que haremos será un diagrama de barras para ver mejor como se distribuyen las diferentes categorías:

Analizaremos los diagramas de alguna de las variables (en total tenemos 53 variables categóricas, no podemos analizarlas todas): *TIPOVIV*, *NACIM*, *SEXO*, *EC* . Los resultados que observemos nos pueden ser de utilidad en el análisis bivariante:



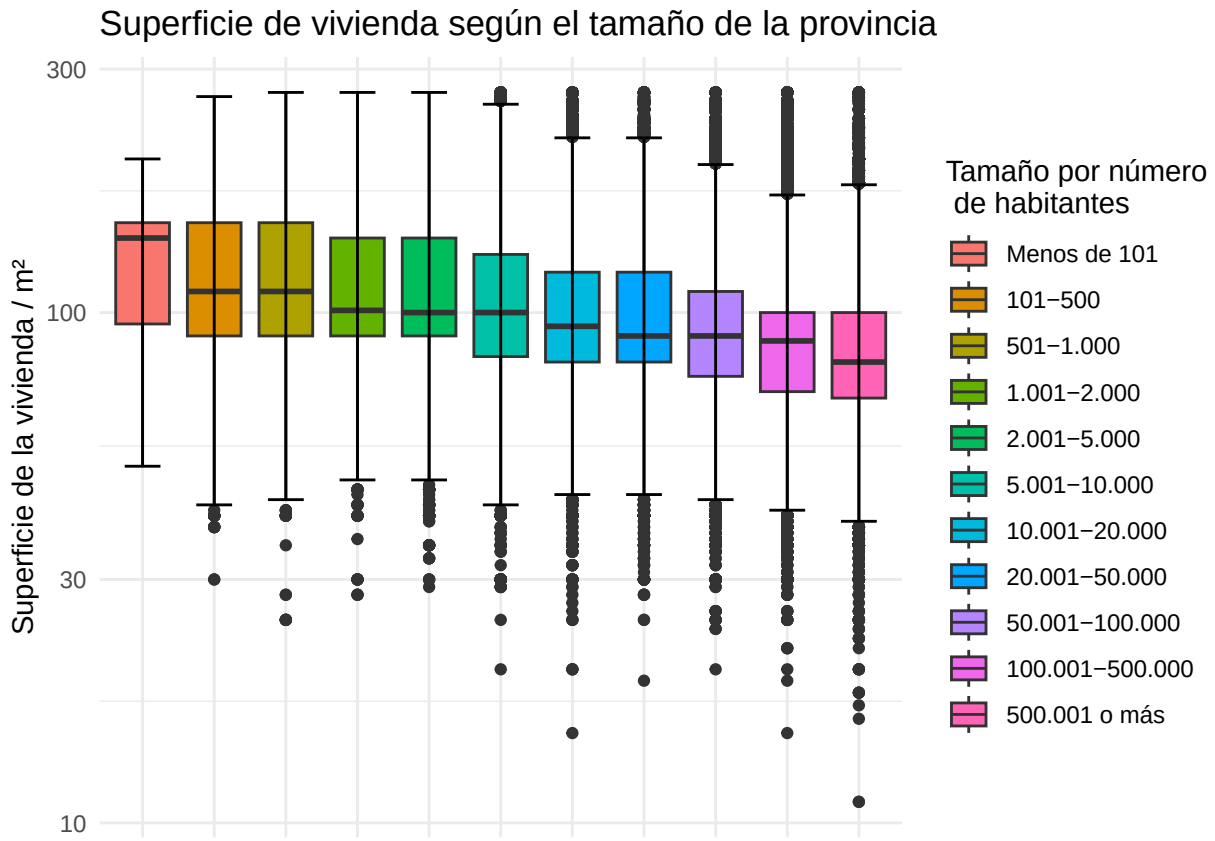
Sobre el tipo de vivienda, observamos claramente que predominan los edificios de 10 viviendas o más, aproximadamente la mitad de las personas viven en este tipo de edificios. Los edificios entre 3 y 9 viviendas y los dos tipos de unifamiliares aparecen en cantidades similares y ya en menos número los edificios de dos viviendas y los edificios para otros usos. Para la variable del lugar de nacimiento, observamos que

de los encuestados la mayoria nacieron en España, casi 25.000 personas nacieron fuera de España de un total de más de 220.000. Observamos paridad entre los encuestado/as, aproximadamente la mitad son hombre y la otra mitad mujeres. Para el estado civil tenemos dos grandes grupos: los casado/as y soltero/as. Aproximadamente cada grupo lo conforman casi 100.000 personas. Las personas restantes pertenecen a divorciado/as, separado/as, viudo/as en proporciones similares.

Análisis bivariante

1. ¿El tamaño de la provincia afecta a los metros de la vivienda?

Para analizar si el tamaño de la provincia tiene alguna relación con los metros de la vivienda usaremos un boxplot. Hemos cambiado la escala a la logarítmica ya que, al haber valores muy grandes, no permitían observar bien el gráfico:




```
## La mediana para las poblaciones de menos de 101 habitantes es 140
```

```
## La mediana para las poblaciones de 500.001 o más habitantes es 82
```

Observando el boxplot vemos como los metros de las viviendas no están muy relacionados con el número de habitantes por m^2, los valores del IQR oscilan entre unos $80m^2$ y unos $150m^2$ de forma aproximada, independientemente del tamaño de la provincia.

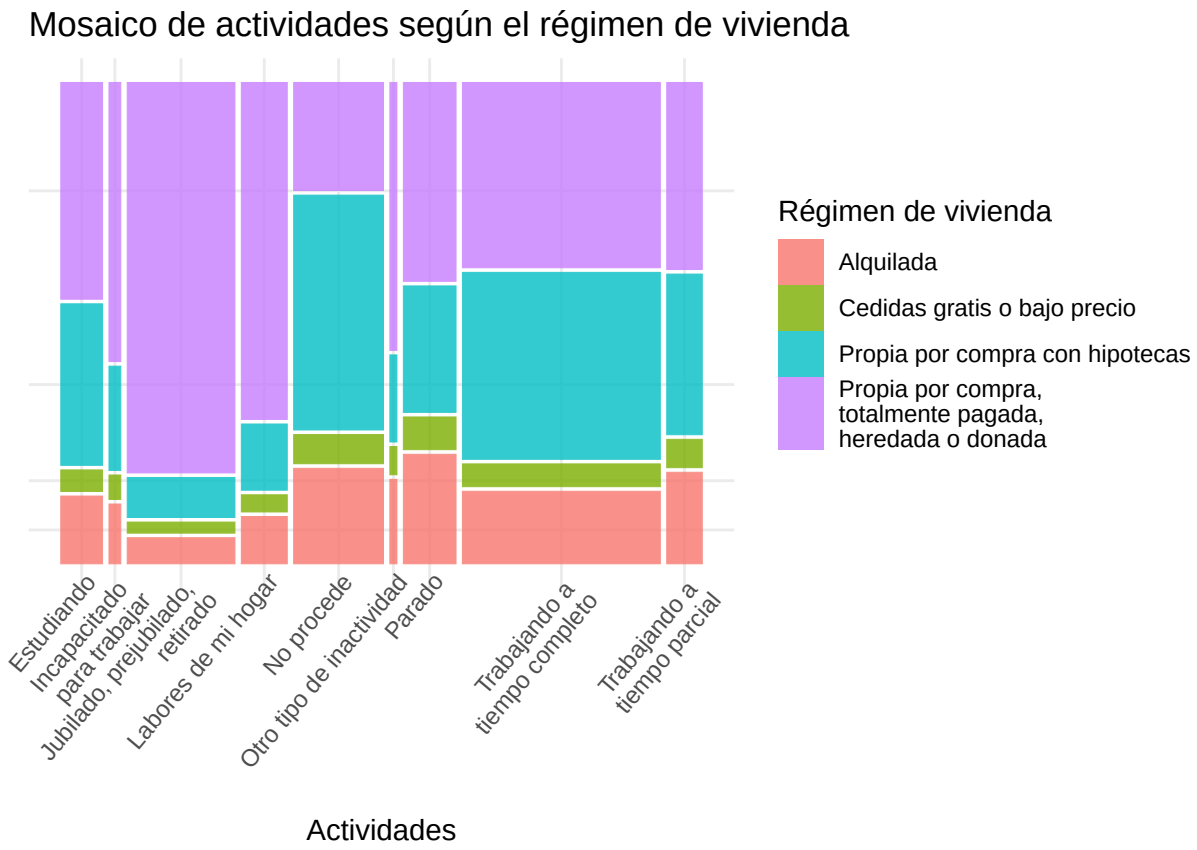
Para los outliers haremos una diferencia: los que se encuentran por debajo de $Q1 - 1.5 \cdot IQR$, y los que se encuentran por encima de $Q3 + 1.5 \cdot IQR$. Para todas las categorías considera, de forma general, outliers las viviendas con menos de $60m^2$. En cambio para los que se encuentran por encima de $Q3 + 1.5 \cdot IQR$ hay una división, para las categorías de más de 20.000 habitantes por m^2 considera outliers las viviendas de más de

$150m^2$ o un poco más ($Q3 + 1.5 \cdot IQR$ es 148 para 500.001 o más habitantes), en cambio para las de menos de 20.000 habitantes por m^2 son las de $250m^2$ aproximadamente ($Q3 + 1.5 \cdot IQR$ es 240 para menos de 101 habitantes). También sobre los outliers, salvo la categoría de "Menos de 101 habitantes", el resto presentan bastantes outliers, y cuanto mayor es el número de habitantes, mayor es el número de outliers.

En último lugar hablaremos sobre la mediana, que analizando el boxplot podemos ver como decrece cuando aumenta el tamaño de la provincia, por lo que cuanto mayor es el número de habitantes por m^2, menos metros suelen tener las vivendas. Sin embargo vemos que esta diferencia no es tan grande, la diferencia se estima en unos $70m^2$.

2. ¿Afecta la actividad que desempeñas con el régimen de la vivienda?

Analizaremos si existe alguna relación entre la actividad que realizan los residentes y el régimen de la vivienda. Como en este caso estamos trabajando con dos variables categóricas, usaremos un mosaico para ver la relación entre estas dos variables:

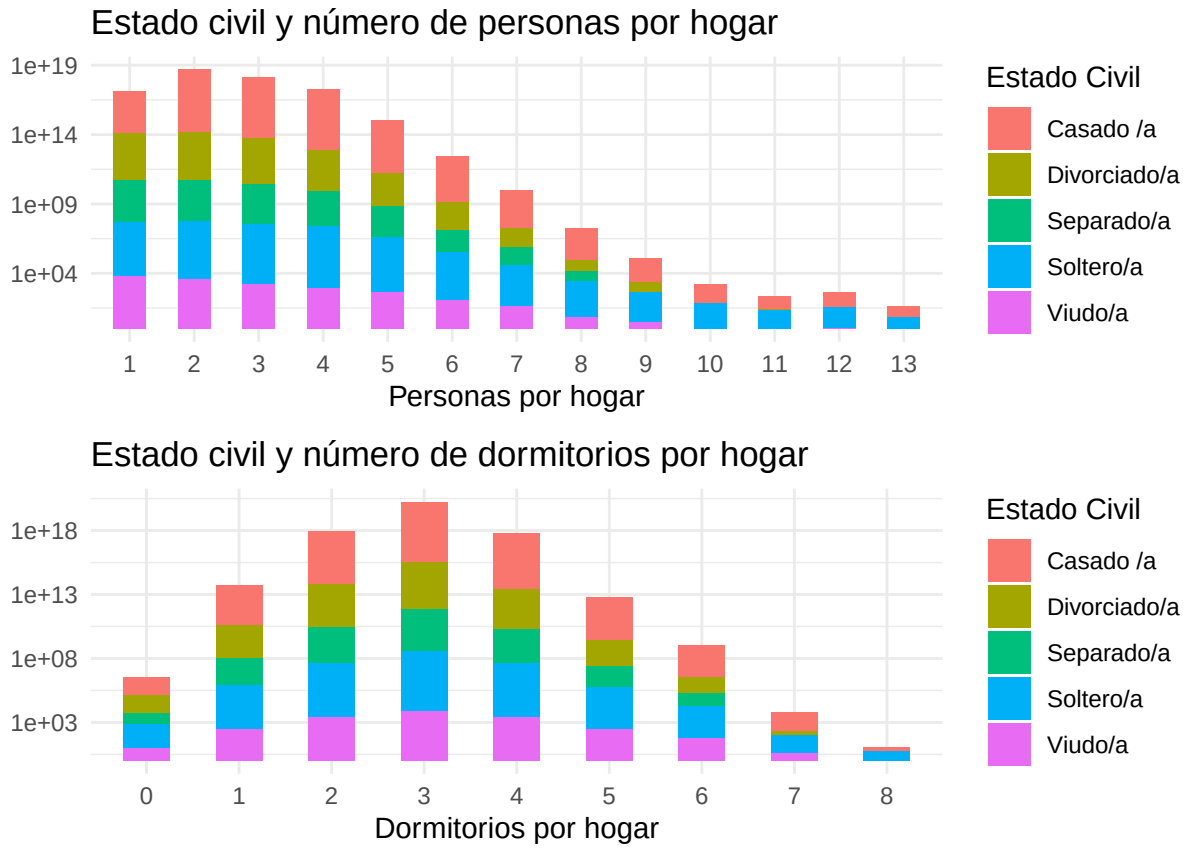


Para que fuera más fácil de comprender esta gráfica, hemos asignado colores según el régimen de vivienda.

A simple vista los casos de casas cedidas gratis o bajo precio por otro hogar, la empresa... ocupan el menor porcentaje de esta variable, independientemente de la actividad que desempeñen los residentes. Podemos destacar que tanto en los residentes jubilados, prejubilados o retirados de una actividad económica previa como en los residentes que desempeñan labores del hogar, con gran diferencia, son los grupos con mayor porcentaje de viviendas propias por compra, totalmente pagada, heredada o donada, ocupando este tipo de régimen de vivienda casi el total de las diferentes categorias. En el caso "no procede" sucede que el mayor porcentaje son las residencias propias por compra con hipoteca, esto se debe a que el tipo de personas que conforman este grupo están formadas por niños y estudiantes parados, entre otros, y cómo seguramente vivan en casas de los adultos que los tutoricen, ese es el motivo de esta distribución. En el resto de actividades, las mayoritarias son las viviendas propias por compra con hipoteca y las viviendas propias por compra, totalmente pagada, heredada o donada, ambas en un porcentaje bastante similar.

3. ¿Tu estado civil tiene relación con tu hogar?

Estudiaremos si el estado civil tiene relación con el número de personas en el hogar y con el número de dormitorios. En este caso también estamos trabajando con dos variables categóricas, y graficaremos unos diagramas de barras:



Para la primera gráfica observamos que, sin importar el número de miembros en un hogar, siempre tenemos representación de soltero/as y casado/as. Esto se puede deber a que hay matrimonios pueden tener bastantes hijos y/o además vivir junto con más familiares, por ello tienen representación hasta para un elevado número de personas por hogar. En el caso de soltero/as, compartir piso puede ser una opción y en algunos casos puede ser con muchas personas, por ello puede estar representado hasta para tantos residentes. Para el resto de estados civiles, viudo/a, separado/a, divorciado/a, tenemos representación hasta para 9 personas por vivienda. Estos 3 casos tienen en común que, almenos, se dejaría de convivir con una persona y ese podría ser el motivo por el cual el número de personas por casa sea menor. Para un número pequeño de residentes observamos que las proporciones entre los diferentes estados civiles son bastante similares, pero conforme aumenta el número de residentes, las proporciones de viudo/a, separado/a, divorciado/a son las que disminuyen más rápidamente.

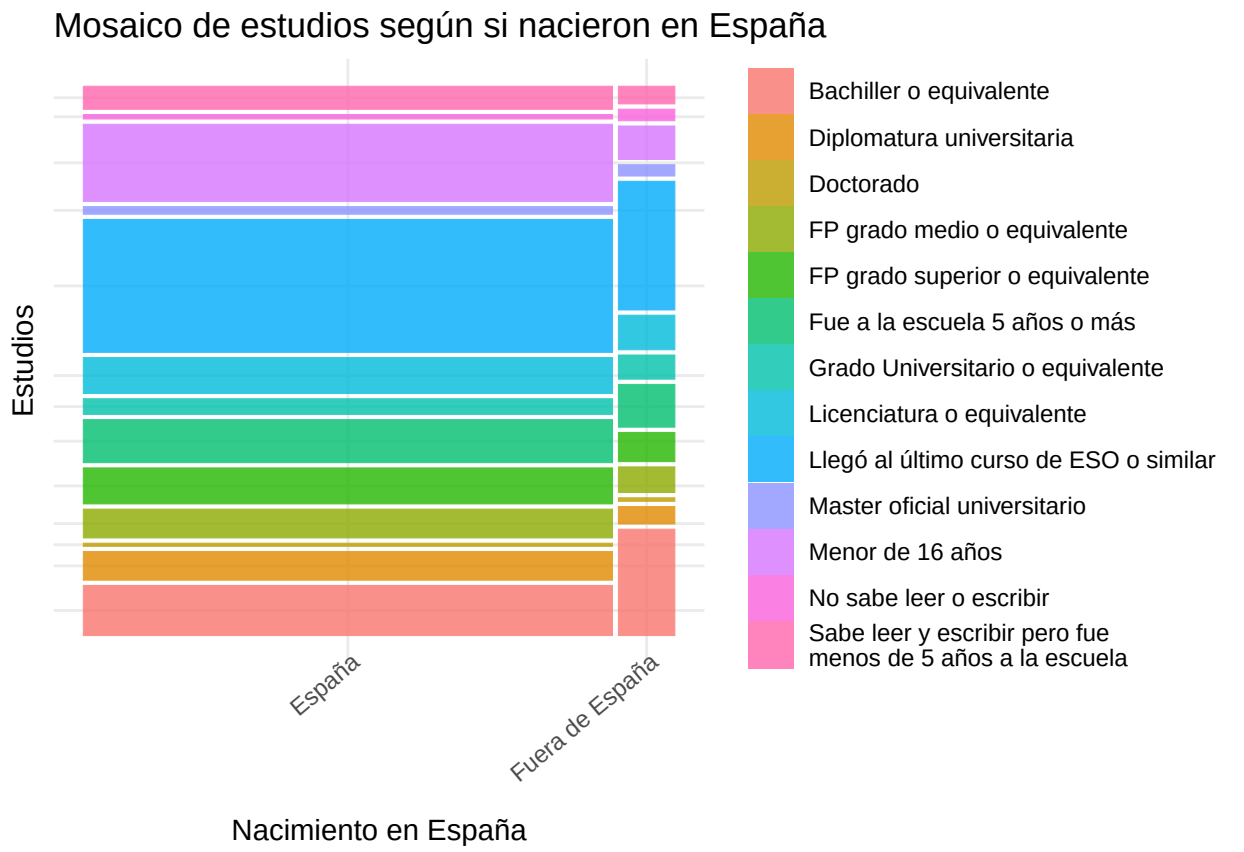
Puede llamarnos la atención que, aunque sea el de menor porcentaje, haya casado/as que vivan solo/as.

Entre 0 y 5 dormitorios vemos que las proporciones entre las diferentes categorias son aproximadamente las mismas, siendo los casado/as y soltero/as los más númerosas. A partir de 6 dormitorios esta predominancia se hace más visibles, desapareciendo algunas categorias como divorciado/as y separado/s para valores más elevados.

4. ¿Existen diferencias entre nacionales y extranjeros en nivel de estudios, ocupación, número de hijos o tipo de vivienda?

Primeramente, trataremos de representar cada variable con respecto a si son nacionales o no. Como para

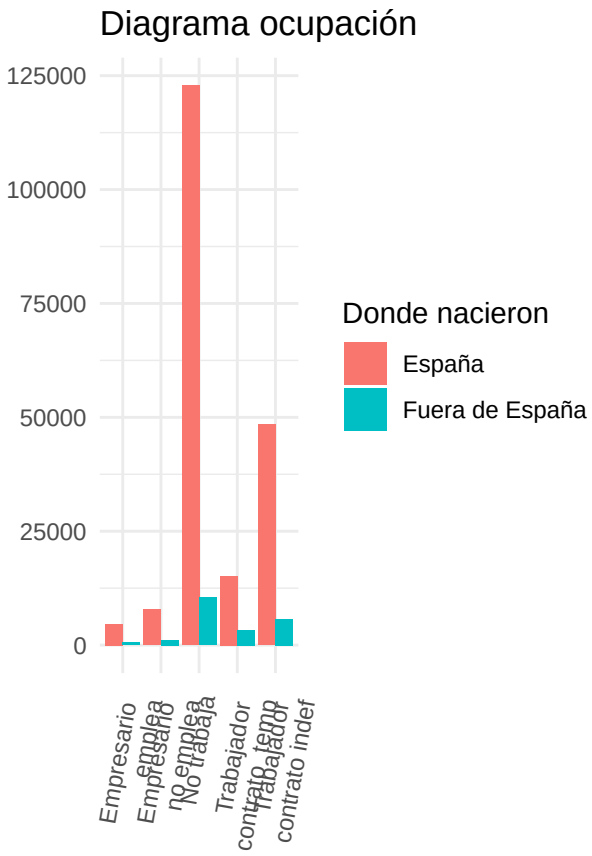
todos los casos ambas variables son categóricas tendremos que optar o por un diagrama de barras o por un mosaico. Para la variable **ESTUDIOS**:

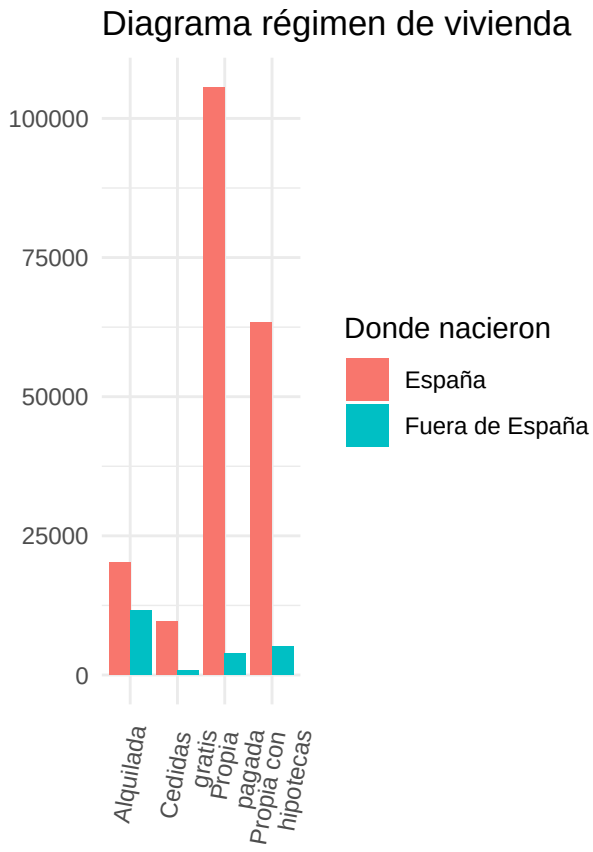


Del mosaico podemos deducir que de las personas que se recopilaron los datos la mayoria nacieron en España. Sobre las personas nacidas en España, el grupo predominante son los que llegaron al último curso de ESO o similar, esto se deberá seguramente a que como bien su nombre indica, Estudios Secundarios Obligatorios, en España son obligatorios. Le sigue el grupo *"Menores de 16 años"*, y aunque se les podría incluir en alguna otra categoria, este estudio ha decidido no incluirlos y considerarlos como casos a parte.

Comparando estos datos con el de las personas nacidas fuera de España, llegar al último curso de ESO o similar sigue siendo el grupo predominante, pero en est caso le sigue de cerca el de las personas que tienen bachiller o equivalentes, cosa que no sucede en el de las personas nacidas en España. El resto de categorias sí que están en proporciones similares tanto para nacido en España como para los que no. Destacar que los que aparecen en menor número son los que no saben leer o escribir, lo que poseen Máster oficial universitario, y los que tienen doctorado. Notamos que estas categorias representan los niveles más bajos y más altos de la educación, la gente que no ha recibido enseñanzas y la que ha alcanzado las titulaciones más elevadas. Tiene sentido que estas sean la de menor número, que pocas personas no sepan ni leer ni escribir es una buena señal del sistema educativo, ya que se pretende que todo el mundo tenga acceso a la educación y de hecho es obligatorio en España. Que pocas personas tengan máster o doctorado también tiene sentido, ya que son estudios que son difíciles de alcanzar y mucha gente pasa antes al mundo laboral.

Hemos recogido en diagramas de barras cómo se distribuyen las variables **OCUPA** y **REGVI** respecto a lugar de nacimiento:





Primeramente analizaremos el diagrama de la ocupación. Sobre las personas nacidas en España, notamos que la mayoría de ellas no trabajan, casi 125.000, siendo más del doble con respecto a la siguiente categoría más numerosa. Recordemos que la categoría "No trabja" está formada por las personas que en la variable **ACTIVIDAD** no Trabajaban ni a tiempo completo ni a tiempo parcial, por lo que estará formada por gente jubilada, en paro, estudiantes, personas con alguna incapacidad para trabajar, que se dediquen a las labores del hogar o otro tipo de inactividad, ese es el motivo de que sea tan numerosa.

Para los nacidos fuera de España, como cuenta con menos muestras, en ninguna de las categorías se superan las 25.000 personas. Sin embargo si destaca que el mayor número se concentra en las que no trabajan, seguido, con cifras similares, los asalariados o trabajadores en ambas categorías. Y con apenas representación quedan los empresarios. Que el mayor número de personas se encuentren las que no trabajan se debe al mismo motivo que para los nacidos en España.

Que en ambas categorías haya pocos empresarios puede explicarse por el hecho de que emprender conlleva ciertos riesgos y requiere un capital que no todos pueden permitirse.

En último lugar analizaremos el régimen de vivienda. Comparando los nacidos en España con los nacidos en el extrangeros notamos una clara diferencia: el primer grupo mayoritáriamente tiene una propiedad comprada pagada totalmente (o heredada o donada), en cambio para el segundo son viviendas alquiladas. Para los nacidos fuera de España una propiedad comprada pagada totalmente (o heredada o donada) pasa a una tercera posición. En ambos casos las casas cedidas gratis o por bajo precio son a las que acceden un menor número de personas y las propiedades compradas con hipoteca ocupan el segundo lugar en sus respectivas categorías, aunque evidentemente, en muy diferente proporción.

Ahora pasaremos a calcular la tabla de contingencia de cada variable respecto a la de lugar de nacimiento y posteriormente calcular la correlación con el test chi-cuadrado.

En los cuatro casos el p-valor es menor que $2.2e-16 < 0.05$, entonces rechazamos la hipotesis nula. En el caso del test chi-cuadrado la hipótesis nula es que no hay asociación entre las variables; son independientes. Al

rechazarla entonces nos quedamos con la alternativa y determinamos que la variables no son independientes.

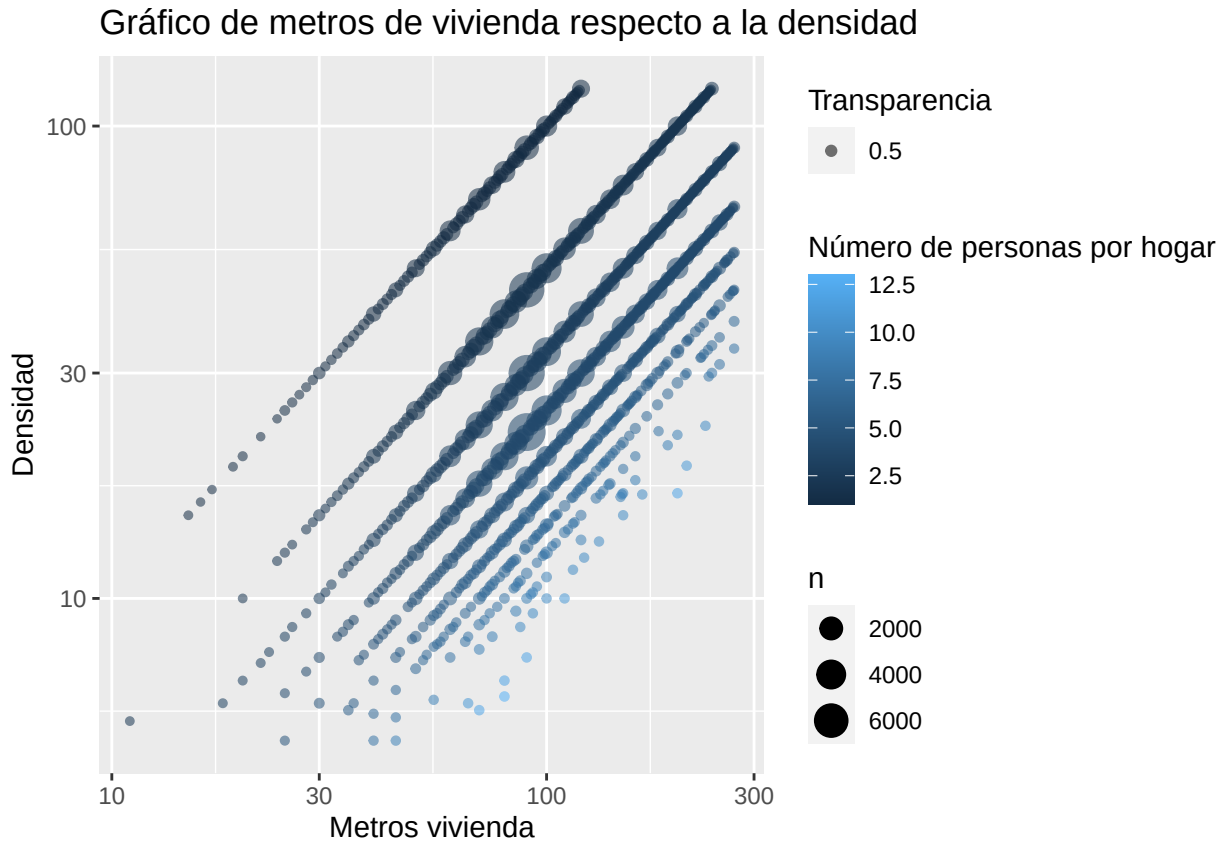
Que el p-valor sea tan pequeño también se puede deber a que la proporción entre nacidos en España y fuera de ella está bastante descompensada.

5. ¿Está correlacionada la densidad de la vivienda con los metros?

```
## La correlación entre la densidad de la vivienda y los metros es  0.6311558
```

Notemos que 0.63 no es un número lo suficientemente grande como para indicar que las variables no están relacionadas linealmente, pero sí nos destaca que están ciertamente relacionadas.

Ahora representaremos los metros de vivienda respecto a la densidad:



Como observación previa, antes de analizar la gráfica, aunque la variable de personas por hogar es categórica, la hemos transformado a numérica para facilitar la comprensión de la gráfica. Al haber tanta cantidad de categorias, impedia entender con claridad el gráfico y de esta forma queda de forma más clara.

Observando la gráfica notamos que existe una cierta relación entre las varibles metros y densidad. Estas dos variables están relacionadas por la variable de personas por hogar, y tiene sentido ya que la variable densidad se obtiene del cociente de estas dos variables restantes.

Ejemplo de uso de los datos

Tras procesar y comprender los datos de nuestro set, ahora queremos explorar las posibilidades que estos datos tienen a la hora de generar contenido informativo. Para ello, crearemos una aplicación con la librería *shiny*, la cual contenga distintas funcionalidades generadaas a partir de los datos.

Aprovechando que las instancias de nuestro dataset contienen la provincia en la cual se realiza la encuesta, un elemento de los que incluiremos será un mapa. Para ello, a través de la API de *geonames*, obtendremos la geolocalización de las distintas provincias que aparecne en el set de datos (en este caso aparecen todas).

Este bloque es el encargado de obtener la latitud y longitud de las distintas provincias (no es necesario ejecutarlo, el dataset está guardado en la carpeta data)

```
provinces <- tab_hogar$Desc[which(tab_hogar$Var == "IDQ_PV")]
get_lat_long <- function(province) {
  aux <- GNsearch(name = province, country = "ES", fcode = "ADM2")
  #Omito el username por privacidad
  result <- c(aux$lat, aux$lng)
  return(result)
}

lat_long <- lapply(provinces, get_lat_long)
lat_long_df <- do.call(rbind, lat_long)
```

Una vez realizada la consulta via API a través de geonames, y obtenidas las latitudes y longitudes de las provincias, arreglamos el dataframe y lo guardamos en la carpeta data para no tener que realizar una consulta a la API cada vez que ejecutemos el código

Ahora que tenemos nuestro dataset con las provincias y sus geolocalizaciones, lo cargamos para trabajar con él. A partir de los datos de nuestrso datasets ya filtrados, obtendremos una serie de representaciones que esperemos nos aporten información valiosa acerca de los mismos.

Todas estas representaciones junto con otros elementos, como varios datasets, las incluimos en una aplicación para que sea más fácil y claro trabajar.