MDPI

Article

Proyecto Análisis Exploratorio de Datos 2023

Javier Hinarejos Albero[1,†,*], Samuel Ortega Mediavilla[1,†,*]

[1] Universitat de València - Escuela Técnica Superior de Ingeniería Avenida de la Universitat s/n 46100 Burjassot. Valencia. España;
* Correspondence: jahial@alumni.uv.es, saorme@alumni.uv.es
† Estos autores contribuyeron de manera equitativa a este trabajo.

Simple Summary: El dataset de variaciones residenciales del INE (2021) contiene los datos de los cambios de residencia con origen y/o destino en España durante el año 2021.

Abstract: Este trabajo presenta un análisis exploratorio de datos basado en un conjunto de datos de variaciones residenciales en España durante el año 2021. Se examinan las fluctuaciones en la residencia de personas, destacando tanto las altas como las bajas ocurridas a lo largo del año. A través de técnicas estadísticas y visualizaciones, se identifican patrones y tendencias en los movimientos de población, proporcionando una comprensión detallada de las dinámicas residenciales en el contexto español durante este periodo. Este análisis contribuye a una mejor comprensión de los cambios demográficos y puede ser fundamental para la toma de decisiones en planificación urbana y políticas de vivienda.

Keywords: AED, ciencia de datos, ine, preprocesamiento, visualización, correlación

Citation: Hinarejos, J.; Ortega, S. Proyecto Análisis Exploratorio de Datos 2023. *Journal Not Specified* **2023**, *1*, 0. https://doi.org/

Received:
Revised:
Accepted:
Published:

1. Datos seleccionados

Hemos escogido los datos de variaciones residenciales en 2021 del INE. Estos datos están disponibles en el siguiente enlace: https://go.uv.es/saorme/ine-var-res-2021.

Adicionalmente, hemos empleado la relación de municipios de 2021, disponible en: https://go.uv.es/saorme/ine-muni-2021.

Se trata de una base de microdatos del INE, procedente de las encuestas realizadas a lo largo de ese mismo año. El dataset contiene información sobre las bajas y altas de residencia, junto a información adicional sobre el encuestado (edad, sexo) y sobre las ubicaciones (provincia, tamaño).

1.1. Preguntas planteadas

- ¿Se concentran las variaciones residenciales durante alguna época concreta del año?
- ¿Hay alguna dependencia de la cantidad de variaciones residenciales con la edad? Si es así, ¿depende también de la edad?
- ¿Cómo es la tasa de migración de España con el extranjero?
- ¿Hay una relación significativa entre el tamaño de los municipios y el número de variaciones que se producen en ellos? ¿Es el éxodo rural un problema actual?

2. Previsualización de los datos

En el fichero principal de datos `md_EVR_2021.txt`, observamos que cada registro contiene una cadena de caracteres de longitud fija. Su interpretación viene detallada en el fichero adicional de metadatos `dr_EVR_2021.xlsx`.

x
601 06106195910801 04202101 111
601 02910199110801 08202101 111
101 03902198210801 06202101 111

Observamos que, como se indicaba, cada entrada tiene la misma longitud, y está codificada en caracteres numéricos cuyo significado hemos de interpretar.

3. Lectura de funciones

El código se ha modularizado para facilitar la comprensión del trabajo realizado. Las funciones definidas para el procesado de este dataset están disponibles en el fichero `ProyectoAED2023_library.R`.

```
Funciones cargadas:
add_comu_variables
apply_dict_to_data
apply_dict_to_variable
check_na_procedence
convert_numeric_vars
create_dict_list
detect_outliers
get_comu_dict
get_country_location
get_dict_from_sheets
get_net_country_movements
get_net_prov_movements
get_prov_location
load_excel_dicts
make_unique_codes
plot_countries_map
plot_residence_variation_map
prepare_aditional_data
prev_objects
rearrange_muni_data
split_raw_data
```

4. Lectura de los ficheros

En primer lugar, leemos la información de todos los ficheros necesarios: el fichero principal, el excel de metadatos y el excel que contiene el diccionario de municipios.

Para ello, hacemos uso de las funciones `readLines()` y `readxl::read_excel()`. Las hojas adicionales del excel de metadatos, que contienen los diccionarios de las distintas variables, se almacenan en una lista.

5. Preprocesado

El preprocesado del dataset comienza con el fichero previsualizado anteriormente, y pretende transformar la información hasta obtener un formato tabular, en este caso un `data.frame`, que sea fácilmente entendible y lo más *tidy* posible.

5.1. Limpieza de entradas de longitud incorrecta

Como primera comprobación, nos aseguramos de que todas las entradas de los datos crudos tienen la longitud adecuada. Este valor está definido en la tabla de metadatos. En este caso, no hay ninguna entrada

5.2. Creación del dataframe

Una vez nos hemos asegurado de que todas las entradas tienen la longitud correcta, dividimos cada una de ellas en distintas variables de un `data.frame` a partir de las posiciones y longitudes definidas también en los metadatos.

5.3. Obtención de los diccionarios

A continuación, extraemos de los ficheros adicionales la información necesaria para interpretar los códigos de los datos crudos. Generamos dos variables:

- `dict_list`: es una lista, en la cual cada elemento corresponde a un diccionario. A su vez, estos elementos son listas, que contienen información útil: la hoja donde se encuentra el diccionario, las variables a las que debe ser aplicado, y un valor lógico que indica si debe ser combinado con el diccionario adicional de municipios.
- `dict_info`: es un `data.frame` en el que almacenamos los códigos y descripciones de todos los diccionarios a aplicar. Dado que los códigos no son únicos entre los distintos diccionarios, también añadimos una columna que indica el nombre del diccionario.

5.4. Interpretación de los códigos de las variables

Una vez obtenidos los diccionarios, los aplicamos a los datos crudos divididos. Para ello, seleccionamos cada uno de los diccionarios descritos en `dict_list` y los aplicamos a todas las variables que ahí se indican, variable a variable.

El intercambio de código a descripción se realiza definiendo cada variable como un `factor` cuyos niveles son los códigos y las etiquetas, las descripciones. Este método ha resultado ser más rápido que la búsqueda de coincidencias variable-código y sustitución con la descripción empleando la función `match()`.

Adicionalmente, verificamos que la interpretación se ha realizado correctamente. Para ello, buscamos los valores no disponibles en el nuevo `data.frame` que no eran `NAs` en el original, y comprobamos si proceden de códigos mal interpretados o, por el contrario, corresponden a entradas en blanco.

```
[1] "check_na_procedence: Omitidas variables sin NAs en la tabla resumen."
```

variable	introduced_na	message
MUNINAC	210261	Todos los NAs corresponden a entradas en blanco.
MUNIALTA	515842	Todos los NAs corresponden a entradas en blanco.
MUNIBAJA	424084	Todos los NAs corresponden a entradas en blanco.
TAMUALTA	452511	Todos los NAs corresponden a entradas en blanco.
TAMUBAJA	662173	Todos los NAs corresponden a entradas en blanco.
TAMUNACI	1544305	Todos los NAs corresponden a entradas en blanco.

El último paso del preprocesado del dataset es la conversión de las variables para al formato adecuado. Está indicado en los metadatos, donde figuran dos tipos:

- N: numérico -> `numeric`
- A: alfanumérico -> `factor`

6. Análisis de las variables

6.1. Resumen del dataset

En primer lugar, realizamos un `summary()` para obtener la información esencial de cada variable.

```
     SEXO                      PROVNAC
 Hombre:1440975   Extranjero       :1544305
 Mujer :1352358   Madrid           : 181183
                  Barcelona        : 168448
                  Valencia/València:  68706
                  Sevilla          :  49597
                  Alicante/Alacant :  37257
                  (Other)          : 743837
      MUNINAC             EDAD
 Marruecos: 197215   Min.   :  0.0
```

```
Madrid   : 142936   1st Qu.: 24.0
Colombia : 140548   Median : 34.0
Venezuela: 101753   Mean   : 35.7
Rumanía  :  99604   3rd Qu.: 47.0
(Other)  :1901016   Max.   :111.0
NA's     : 210261
    MESNAC          ANONAC
Min.   : 1.00   Min.   :1909
1st Qu.: 3.00   1st Qu.:1973
Median : 6.00   Median :1987
Mean   : 6.44   Mean   :1985
3rd Qu.: 9.00   3rd Qu.:1997
Max.   :12.00   Max.   :2021

       CNAC                      PROVALTA
España   :1397563   Extranjero       : 452511
Marruecos: 192328   Madrid           : 369788
Colombia : 117941   Barcelona        : 347321
Rumanía  : 109781   Valencia/València: 142855
Venezuela:  69375   Alicante/Alacant : 115759
Italia   :  68134   Málaga           : 100499
(Other)  : 838211   (Other)          :1264600
             MUNIALTA          MESVAR
No Consta         : 221711   Min.   : 1.00
Madrid            : 164977   1st Qu.: 4.00
Baja por Caducidad: 135433   Median : 7.00
Barcelona         : 106651   Mean   : 6.68
València          :  39015   3rd Qu.:10.00
(Other)           :1609704   Max.   :12.00
NA's              : 515842
    ANOVAR                  PROVBAJA
Min.   :2021   Extranjero       : 662173
1st Qu.:2021   Madrid           : 354922
Median :2021   Barcelona        : 331140
Mean   :2021   Valencia/València: 121356
3rd Qu.:2021   Alicante/Alacant :  97266
Max.   :2021   Málaga           :  77088
               (Other)          :1149388
     MUNIBAJA
Madrid   : 172651
No Consta: 135696
Barcelona: 114877
Marruecos:  53873
Colombia :  49732
(Other)  :1842420
NA's     : 424084
                                          TAMUALTA
Municipio no capital hasta 10.000 habitantes:515842
Municipio no capital de 10.001 a 20.000     :250486
Municipio no capital de 20.001 a 50.000     :404947
Municipio no capital de 50.001 a 100.000    :293254
Municipio no capital  de más de 100.000     :220249
Municipio capital de provincia              :656044
NA's                                        :452511
```

```
                                          TAMUBAJA
Municipio no capital hasta 10.000 habitantes:424084
Municipio no capital de 10.001 a 20.000     :217365
Municipio no capital de 20.001 a 50.000     :352103
Municipio no capital de 50.001 a 100.000    :249767
Municipio no capital  de más de 100.000     :204544
Municipio capital de provincia              :683297
NA's                                        :662173
                                         TAMUNACI
Municipio no capital hasta 10.000 habitantes: 210261
Municipio no capital de 10.001 a 20.000     :  81725
Municipio no capital de 20.001 a 50.000     : 131825
Municipio no capital de 50.001 a 100.000    : 102559
Municipio no capital  de más de 100.000     : 104088
Municipio capital de provincia              : 618570
NA's                                        :1544305
```

Seguidamente, observamos el tipo de cada variable para confirmar que estén en el formato adecuado empleando la función `str()`.

```
'data.frame':   2793333 obs. of  16 variables:
 $ SEXO    : Factor w/ 2 levels "Hombre","Mujer": 2 2 1 1 2 1 1 1 2 2 ...
 $ PROVNAC : Factor w/ 53 levels "Araba/Álava",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ MUNINAC : Factor w/ 8316 levels "Albania","Austria",..: NA NA NA NA NA NA NA NA
 $ EDAD    : num  61 29 39 25 25 19 15 12 29 11 ...
 $ MESNAC  : num  6 10 2 8 10 7 10 1 10 1 ...
 $ ANONAC  : num  1959 1991 1982 1995 1996 ...
 $ CNAC    : Factor w/ 202 levels "Albania","Austria",..: 7 7 7 7 7 7 7 7 7 7 ...
 $ PROVALTA: Factor w/ 53 levels "Araba/Álava",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ MUNIALTA: Factor w/ 8316 levels "Albania","Austria",..: NA NA NA 250 250 250 25
 $ MESVAR  : num  4 8 6 3 12 8 11 12 3 3 ...
 $ ANOVAR  : num  2021 2021 2021 2021 2021 ...
 $ PROVBAJA: Factor w/ 53 levels "Araba/Álava",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ MUNIBAJA: Factor w/ 8316 levels "Albania","Austria",..: NA NA NA NA NA NA NA NA
 $ TAMUALTA: Factor w/ 6 levels "Municipio no capital hasta 10.000 habitantes",..:
 $ TAMUBAJA: Factor w/ 6 levels "Municipio no capital hasta 10.000 habitantes",..:
 $ TAMUNACI: Factor w/ 6 levels "Municipio no capital hasta 10.000 habitantes",..:
```

Tras observar las diferentes variables del conjunto de datos, decidimos eliminar aquellas variables que no aportan información valiosa en nuestro conjunto de datos:

- **MESNAC**: El mes de nacimiento del encuestado no es relevante.
- **ANOVAR**: Todos los datos provienen del año 2021.

Por otra parte, creamos nuevas variables que contienen las comunidades autónomas, para poder analizar también las variaciones residenciales entre ellas. Para ello, aprovechamos el diccionario de municipios, ya que en él también aparecen codificadas las comunidades autónomas, y nos permite relacionar las provincias con las comunidades. Esta variable es más interesante que las provincias o los municipios, ya que en estás dos el número de categorías es muy elevado.

En las nuevas variables de comunidades autónomas, las localizaciones en el extranjero están codificadas como "Extranjero".

*6.2. Valores faltantes (*NA*)*

En los resúmenes mostrados podemos observamos que las variables de municipio y tamaño contienen numerosos valores faltantes. Además, en algunas de las variables

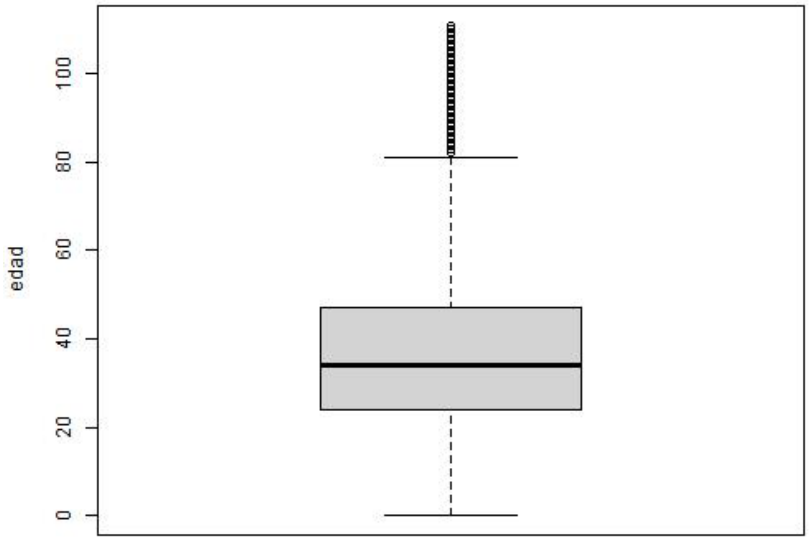
categóricas, aparecen niveles cuyo significado es equivalente a un dato faltante de cara al análisis: “No Consta”, “(Other)”, “Baja por Caducidad”.

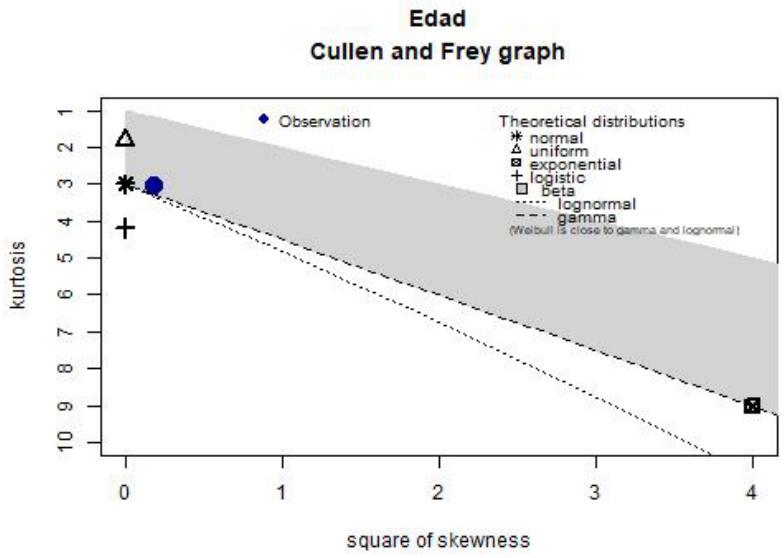
En ninguno de los casos es posible realizar una imputación de datos faltantes, ya que no disponemos de ninguna información que nos permita obtener el municipio de alta o de baja faltante.

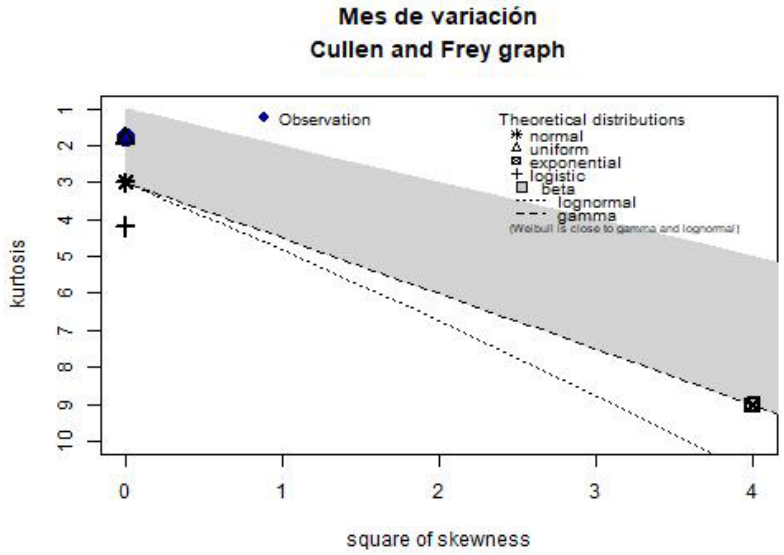
Encontramos un máximo de municipios faltantes, tanto de alta como de baja, en la provincia de Barcelona. Por ello, podríamos afirmar que estos resultados son, en parte, *MAR*, aunque también hay un número considerable de valores faltantes correspondientes en el resto de provincias, lo que indica más bien que es *MCAR*.

Por otra parte, los valores faltantes de las variables de tamaño de municipio están relacionadas en su totalidad a entradas en las que el municipio correspondiente (ALTA, BAJA o NAC) es un país extranjero. Por ello, estos valores faltantes son *MAR*.

6.3. Análisis univariante



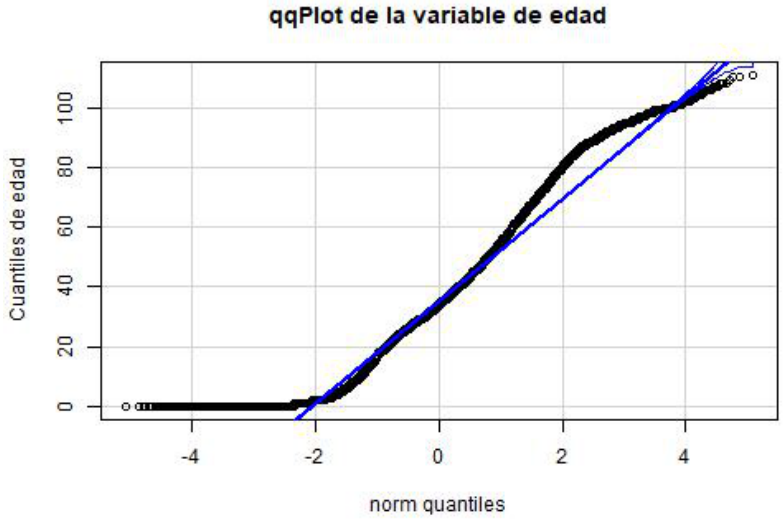




Mediante los gráficos de *Cullen* y *Frey* observamos que la variable **MESVAR** la podemos aproximar mediante una función uniforme. Por tanto, podemos eliminarla también ya que no aporta valor a nuestro conjunto de datos. Un posible añadido sería cargar datos de años diferentes y estudiar la serie temporal para observar si hay alguna relación o estacionalidad entre mudarse y el mes de cambio de residencia.

Por otra parte, la variable **EDAD** nos indica que se puede ajustar bajo una distribución gamma. Esto es coherente con la distribución, ya que la mediana es 35.7 y sin embargo, alcanza valores de hasta 111 años, por lo que es considerablemente asimétrica. Además, es posible explicar el valor obtenido para la mediana de edad: gran parte de la población logra la estabilidad ecónomica y/o familiar en la treintena, por lo que es alrededor de esta edad en la cual hay más mudanzas, y por tanto se registran más variaciones residenciales.

Podemos realizar un análisis adicional de la variable **EDAD**. A pesar de ser una distribución gamma, estudiamos cuánto se aproxima a una distribución normal mediante la función `qqPlot()`.



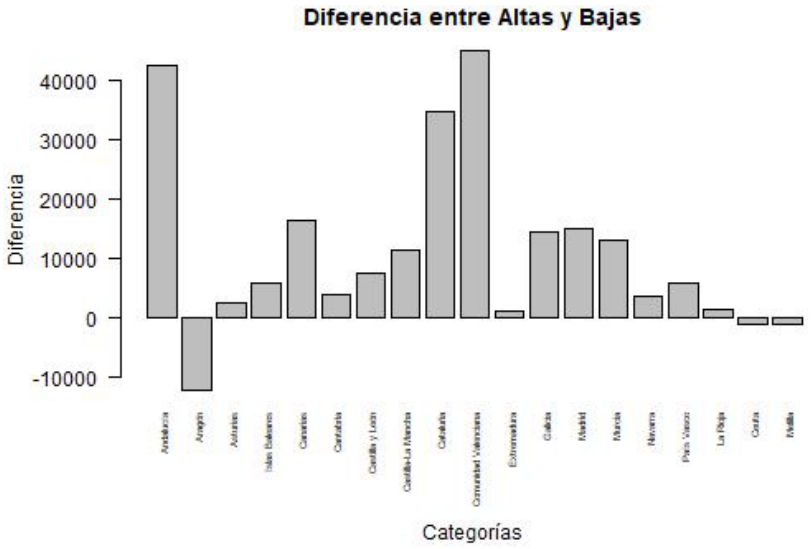
Observamos que no sería una representación idónea, ya que la cola inferior se separa considerablemente de los cuantiles normales. Esto justifica el gran número de *outliers* detectados anteriormente en el *boxplot*, ya que esta función supone que la distribución que grafica es normal.

6.4. Análisis univariante (Variables Categóricas)

Tras analizar las edades y el mes de variación, planteamos otra cuestión de vital importancia como es el hecho de estudiar el éxodo rural. Para ello, decidimos estudiar el movimiento entre provincias.

El dato más significativo que se observa es que España es un país con un mayor número de inmigrantes que de emigrantes y por consecuencia, la población en las diferentes provincias españolas aumenta. Además, otro dato curioso es que no se observa un decrecimiento en las provincias del interior de España "La España despoblada".

Como hemos comentado anteriormente, si analizamos todas las provincias o municipios tenemos muchos niveles dentro del factor. Por ello, vamos a intentar obtener información más relevante a través del estudio de las comunidades autónomas.

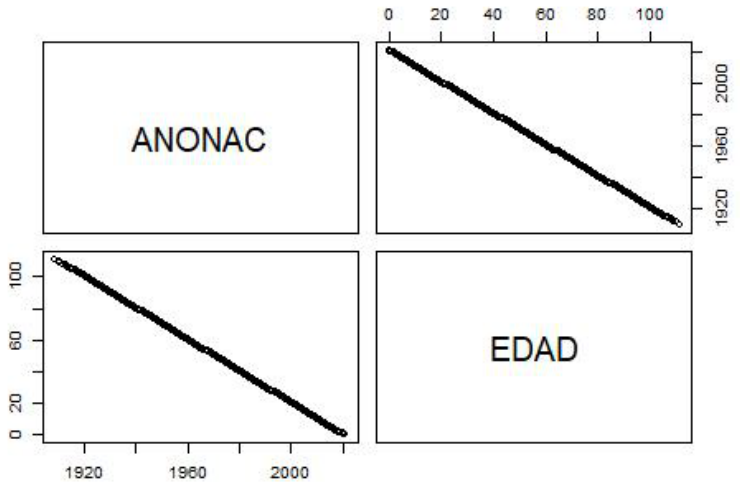


En esta gráfica podemos ver que únicamente las comunidades de Aragón, Ceuta y Melilla presentan variaciones netas negativas, y La Comunidad Valenciana y Andalucía son las comunidades que cuentan con un mayor incremento.

6.5. Análisis bivariante

6.5.1. Numérica - Numérica

La variable **ANONAC** debería tener una gran correlación con la variable **EDAD**.



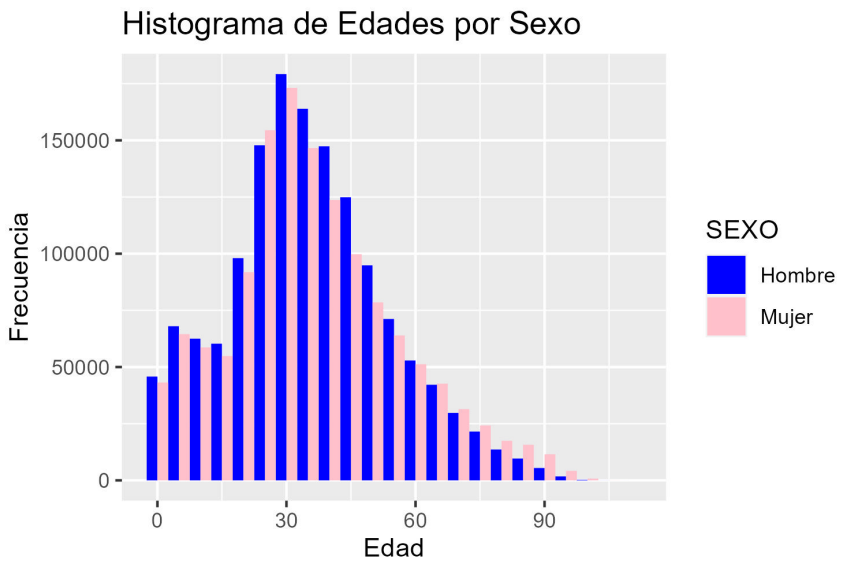
Efectivamente, como era de suponer, obtenemos que las dos variables tienen un alto grado de correlación negativa, ya que $EDAD = 2021 - ANONAC$.

```
       ANONAC EDAD
ANONAC    372 -371
EDAD     -371  372

       ANONAC EDAD
ANONAC      1   -1
EDAD       -1    1

       ANONAC EDAD
ANONAC      1   -1
EDAD       -1    1
```

La correlación de Pearson sirve para cuando la relación entre dos variables es lineal, mientras que la de Spearman es robusta frente a relaciones no lineales en datos ordenados. Por esta razón, ambas correlaciones son iguales, ya que la relación entre las variables es lineal.

6.5.2. Numéricas- Categóricas

Seguidamente, utilizando la librería `ggplot` vamos a representar un histograma de edades por sexo, ya que queremos conocer la edad a la cual la gente suele cambiar de residencia y si existe alguna diferencia significativa entre hombres y mujeres a la hora de tomar esta decisión.



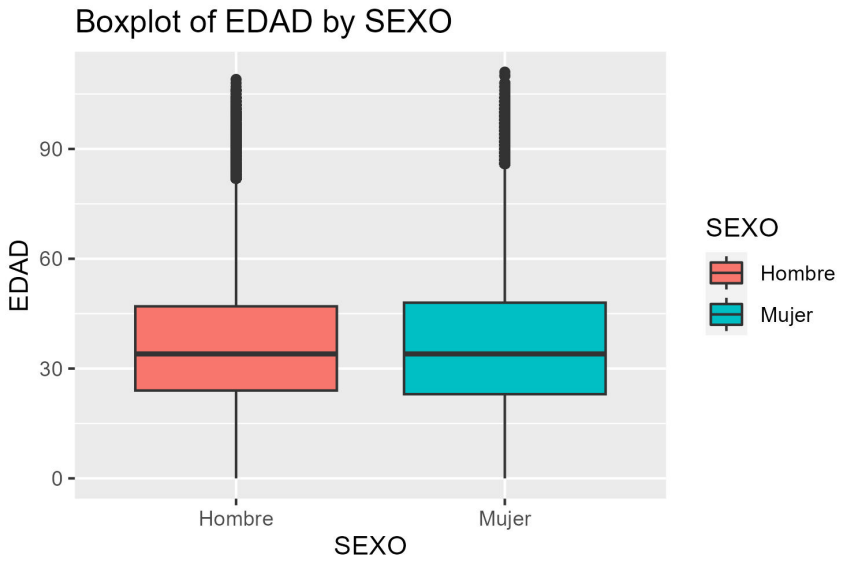
En esta gráfica, se observa que la gente cambia más de residencia alrededor de los 30 años, lo cual tiene sentido, ya que es la etapa de la vida en la que muchas personas deciden independizarse o formar una nueva familia.

Si queremos estudiar ambas colas, un detalle importante a tener en cuenta y que está apoyado científicamente es que las mujeres viven más de media que los hombres y lo podemos observar a edades tardías, ya que hay una diferencia significativa entre hombres y mujeres a esa edad.

Por otra parte, también se observa un pequeño máximo local, que puede estar explicado por el hecho de que tras el nacimiento de los hijos o, en muchas ocasiones, del segundo hijo, las familias suelen tomar la decisión de mudarse a un hogar más amplio.

Seguidamente, mediante un test T veremos si podemos considerar que las medias para hombres y mujeres son iguales.

```
    Welch Two Sample t-test

data:  mujeres$EDAD and hombres$EDAD
t = 28, df = 3e+06, p-value <2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.599 0.689
sample estimates:
mean of x mean of y
     36.1      35.4
```

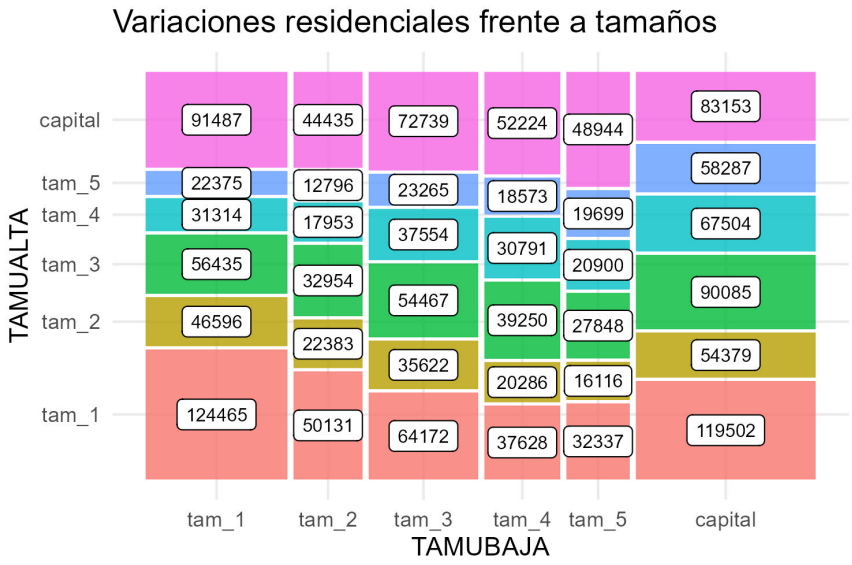


Por tanto, rechazamos la hipótesis nula de que las medias de edad de los hombres y las mujeres es la misma.

6.5.3. Categóricas - Categóricas

Para seguir con el estudio del éxodo rural, podemos representar la relación entre el tamaño de los municipios de alta y de baja en un mosaico. Por limpieza, hemos recodificado las categorías de tamaño de la siguiente manera:

- tam_1: Municipio no capital hasta 10.000 habitantes
- tam_2: Municipio no capital de 10.001 a 20.000
- tam_3: Municipio no capital de 20.001 a 50.000
- tam_4: Municipio no capital de 50.001 a 100.000
- tam_5: Municipio no capital de más de 100.000
- capital: Municipio capital de provincia



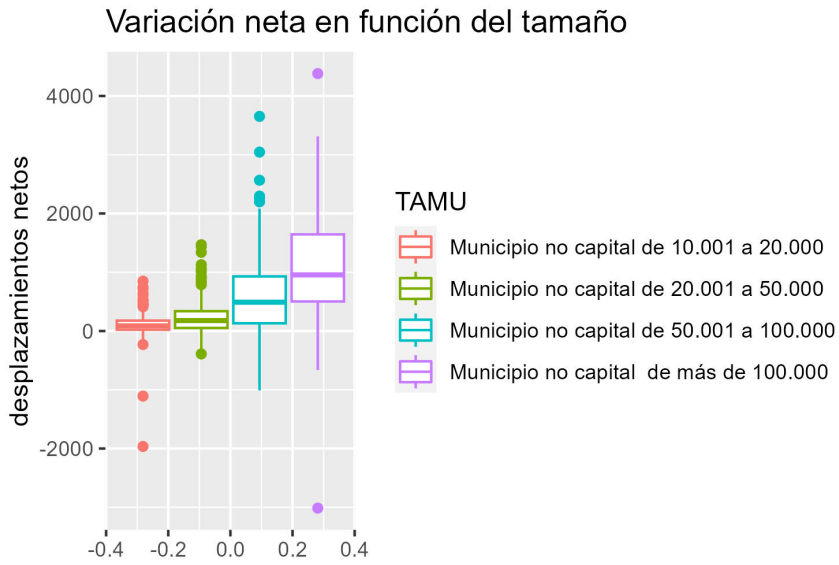
No observamos que haya una clara diferencia. Cabe destacar que el máximo se da para movimientos entre municipios de menos de 10000 habitantes, lo cual es esperable ya que representan la categoría con mayor número de municipios.

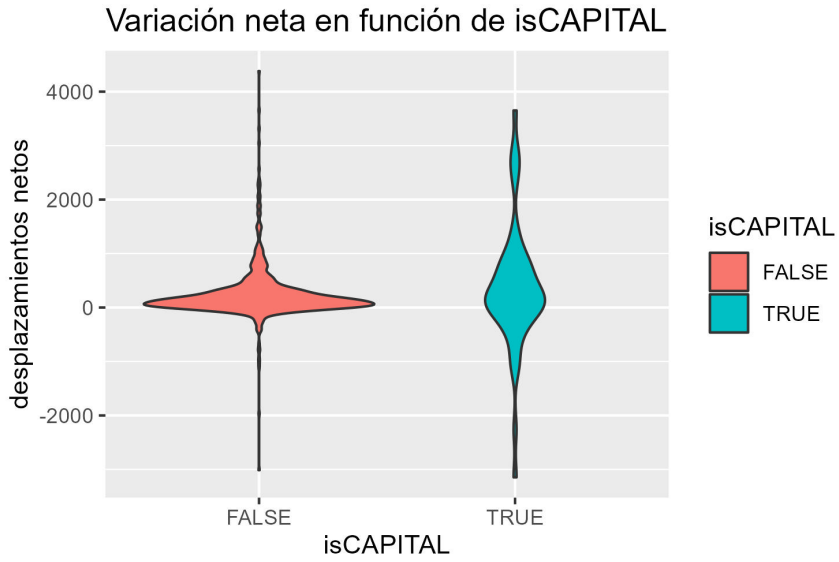
Para complementar este análisis, transformamos nuestros datos a fin de obtener un `data.frame` con la siguiente estructura:

- **MUNI**: contiene todos los valores únicos de las variables **MUNIALTA** y **MUNIBAJA**.
- **TAMU**: valor correspondiente de **TAMUALTA** / **TAMUBAJA**.
- **isCAPITAL**: valor lógico que indica si el municipio es capital.
- **EDAD**: media de la edad de los desplazados desde ó hasta cada municipio.
- **MES**: moda del mes en el que se producen los movimientos desde ó hasta cada municipio.
- **nBAJAS**: número de bajas en cada municipio.
- **nALTAS**: número de bajas en cada municipio.

Las variables adicionales **nTOTAL** y **nNETO** son la suma y la diferencia de las últimas dos variables listadas.

Ahora podemos realizar diferentes representaciones sobre este nuevo dataset transformado.





Como cabía esperar, el número las variaciones residenciales netas de los municipios más grandes es más elevado, esto es, sí observamos un cierto grado de centralización por el cual un gran número de personas se desplaza hacia las ciudades más grandes.

Por otra parte, observamos que hay una mayor cantidad de municipios no capitales con variaciones netas de menor valor, mientras que en las capitales se distribuyen en mayor medida hacia valores más extremos. Debemos aclarar que, para mejorar la calidad del gráfico, se ha omitido el dato correspondiente a Zaragoza, capital de provincia, que es -16762. Este número es muy extremo en comparación al resto.

6.6. Análisis interactivo: mapas

Empleamos la librería `leaflet` para crear mapas interactivos sobre los que representamos algunos de los resultados obtenidos en el análisis. También hemos usado la librería ggmap https://journal.r-project.org/archive/2013-1/kahle-wickham.pdf) para obtener las longitudes y latitudes de las distintas ubicaciones. En este documento se expone una imagen fija de uno de ellos. Para poder consultar los mapas en su totalidad, ejecute las celdas de este apartado en el documento `ProyectoAED2023.Rmd`.

6.6.1. Características

Se usa el test Chi-cuadrado. Este test supone una hipótesis de partida H_0 (Son independientes) y dependiendo del resultado del test, se acepta o no:

p<0.05: Rechazamos hipótesis p $\geq$ 0.05: Aceptamos Ho

```
    Pearson's Chi-squared test

data:  tablacontingencia1
X-squared = 1e+07, df = 361, p-value <2e-16
```

Por tanto, como p $\geq$ < 0.05, rechazamos la hipótesis nula. y por tanto, concluimos que las variables COMUBAJA y COMUALTA son dependendientes.

6.7. Análisis de outliers

Respecto a los outliers, únicamente hemos analizamos la variable numérica **EDAD**. Los outliers detectados corresponden a valores anómalos en comparación a la distribución teórica, sin embargo, en ningún caso son valores imposibles de edad, como serían valores negativos o extremadamente elevados.

```
     method       n nMiss   nOut lowLim upLim
1 percentil 2793333     0 254576    4.0  71.0
2 tresSigma 2793333     0   5058  -22.1  93.5
3    hampel 2793333     0  23973  -19.4  87.4
4   boxplot 2793333     0  54238  -10.5  81.5
  minNom maxNom
1      5     70
2      0     93
3      0     87
4      0     81
```

La regla del identificador de Hampel es la única que no considera que la distribución sea gaussiana. Sin embargo, también etiqueta como outliers valores que realmente son posibles, ya que su límite superior es 87.4. Por otra parte, pese a que la edad no sigue una distribución gaussiana, la regla 3 sigma es muy poco agresiva para la detección de outliers y por tanto, es la que menos valores detecta como anómalos.

También realizamos un estudio de los outliers comparando ambos sexos: la mediana de las distribuciones es la misma, sin embargo, la media es mayor para las mujeres que para los hombres debido a la cola superior de las mujeres, razonada ya anteriormente a partir de su mayor esperanza de vida.

7. Conclusiones finales

Podemos afirmar que, durante el año 2021 en España, las variaciones residenciales no presentaron ninguna dependencia significativa con la época del año. Por otra parte, sí observamos que la edad es un factor importante en cuanto al cambio de residencia: hay un máximo absoluto en torno a los 35 años, que es coherente con la edad de independización definitiva de muchas familias. También concluimos que esta relación no es independiente del sexo.

La tasa de migración de España con el extranjero durante el año estudiado es positiva y de valor elevado. Sin embargo, este resultado puede estar influido por la amplia cantidad de variaciones de salida codificadas como "Baja por Caducidad".

Por último, observamos que gran cantidad de las variaciones residenciales se producen entre municipios pequeños, pero no suponen un desplazamiento neto elevado. No obstante, en las ciudades más grandes la tasa neta de variaciones residenciales es mucho más grande, lo cual es un indicio del proceso de centralización actual.

En conclusión, hemos logrado importar, procesar, interpretar y analizar el dataset propuesto. Para ello, hemos hecho uso de numerosas funciones y librerías que nos han permitido realizar este proyecto de manera eficiente y obtener las conclusiones descritas.

Supplementary Materials: No hay información de apoyo disponible.

Author Contributions: S.O. y J.H. hicieron la búsqueda y selección del dataset; S.O. realizó el preprocesamiento de los datos; J.H. realizó un análisis estadístico profundo de los datos procesados; S.O. y J.H. realizaron las representaciones gráficas; S.O. y J.H. redactaron el trabajo.

Funding: Este proyecto no ha recibido financiación externa.

Institutional Review Board Statement: El estudio se ha realizado de acuerdo a la licencia de libre disposición de los datos anonimizados del INE.

Informed Consent Statement: No aplicable.

Data Availability Statement: Los resultados de este proyecto se pueden encontrar en el repositorio de GitHub creado a fin de contenerlo: https://github.com/esedesam/ProyectoAED2023.git.

Acknowledgments: Hasta la fecha de publicación, no se ha recibido ningún tipo de financiamiento para este proyecto.

Conflicts of Interest: Los autores declaran la ausencia de conflictos de intereses.

Sample Availability: Los datos están disponibles en la página web del INE.

Abbreviations

The following abbreviations are used in this manuscript:

INE	Instituto Nacional de Estadística
AED	Análisis Exploratorio de Datos
MCAR	Missing Completely At Random
MAR	Missing At Random
NA	Not Available