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: 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.

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

3. Lectura de funciones

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_variabl
```

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

5.1. Limpieza 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.

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 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. Análisis univariante

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.

```
'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 ...
$ 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 ...
$ 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.

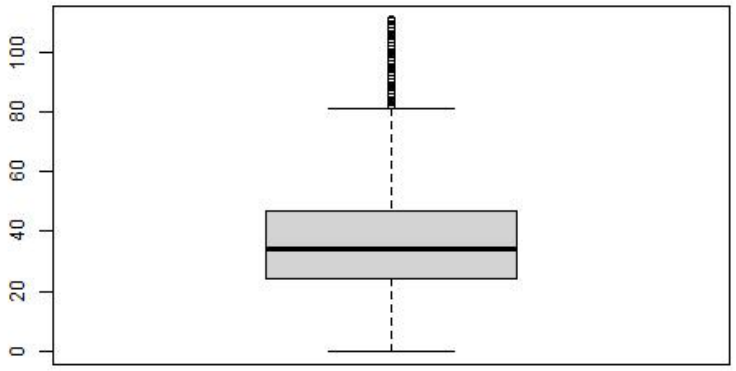
Decidimos eliminar las variables:

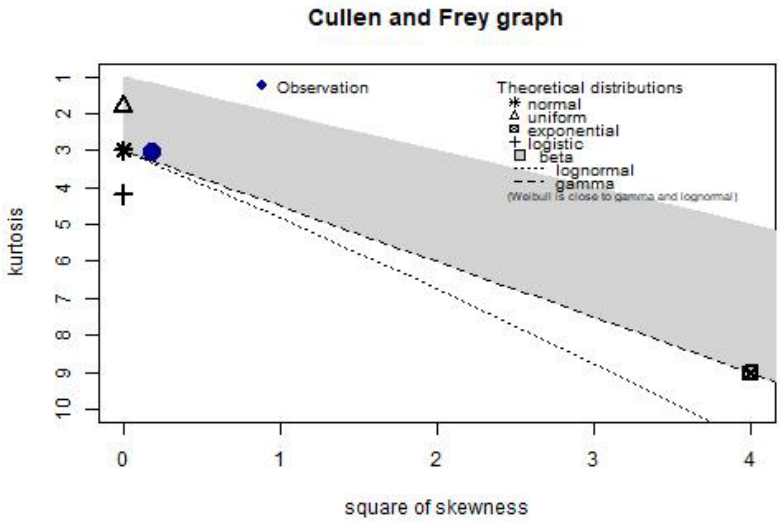
- **MESNAC**: Consideramos que no aporta valor el mes de nacimiento de una persona.
- **ANOVAR**: Todos los datos provienen del año 2021.

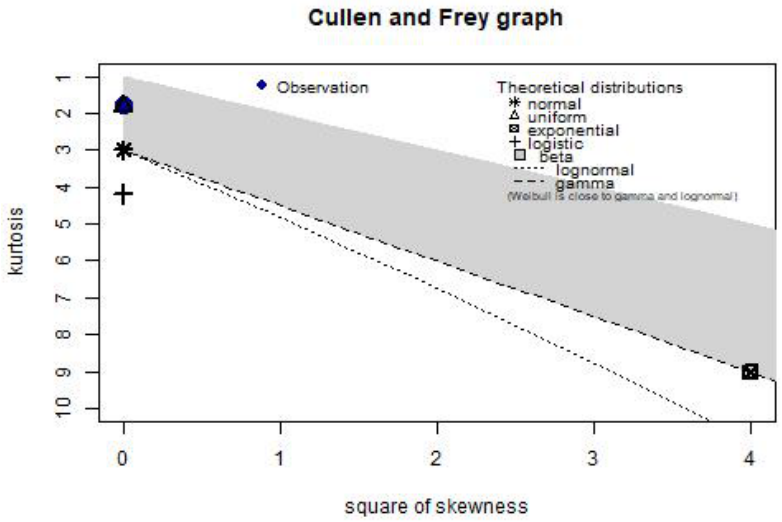
Otra variable que es posible que no sea de mucho interés es **MESVAR**, pero de momento decidimos no eliminarla para estudiar si existe alguna época del año en la cual las personas decidan cambiar de residencia con mayor frecuencia.

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. Dado que las variables de municipios presentan muchos NAs, intentamos imputar las comunidades autónomas faltantes buscando coincidencias de la provincia.

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

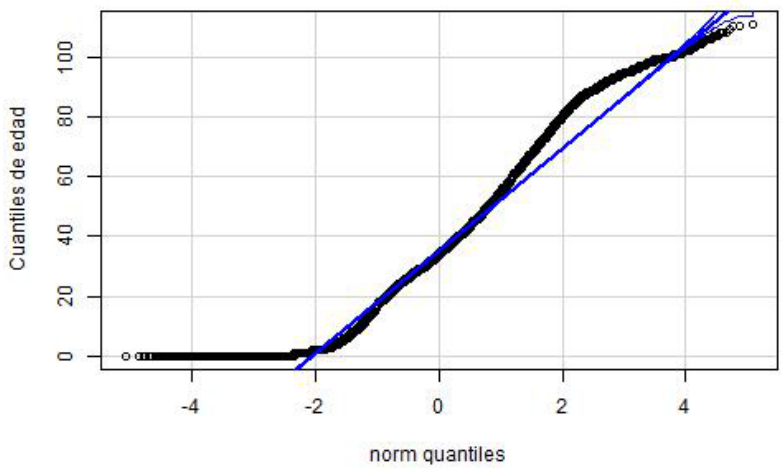






Mediante los gráficos de Cullen y Frey observamos que la variable MESVAR la podemos aproximar mediante una función uniforme. Por tanto, podemos eliminarla también ya que no aporta valor a nuestro conjunto de datos. Si tuviésemos datos de más años podríamos calcular la serie temporal y observar si hay alguna relación 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, lo cual tiene sentido, ya que la mediana es 35.7 y sin embargo, alcanza valores de hasta 111 años. Esto es debido a que la mayor parte de la gente se muda alrededor de la treintena y una vez ya han conseguido un trabajo y una familia estable, la decisión de cambiar de residencia es cada vez más complicada.

A pesar de ser una distribución gamma, mediante la función `qqPlot()` vemos si se puede aproximar también mediante una distribución normal y observamos que no sería una representación idónea.



6.2. Análisis univariante (Variables Categóricas)

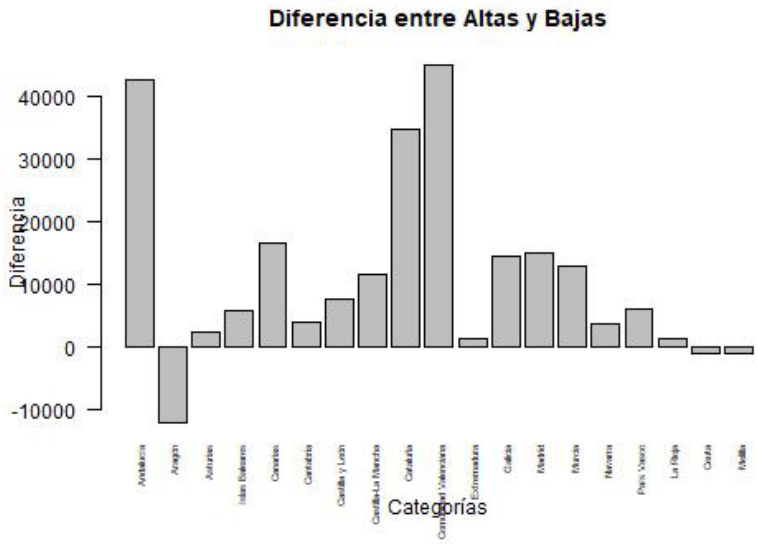
Tras analizar las edades y el mes de variación, planteamos otra cuestión de vital importancia como es el hecho de estudiar el éxodo rural. Para ello, decidimos estudiar el movimiento entre provincias.

```
[1] TRUE
```

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".

Si analizamos todas las provincias o municipios, tenemos muchos niveles dentro del factor, así que vamos a intentar obtener información más relevante a través del estudio de las comunidades autónomas.

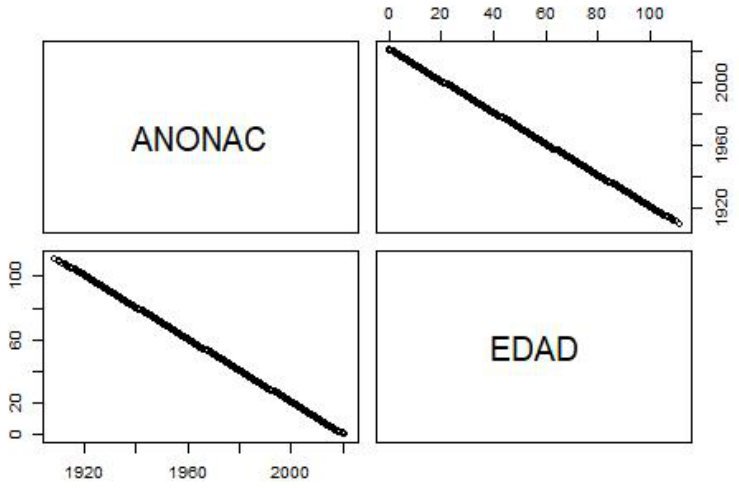
```
[1] TRUE
```



6.3. Análisis bivariante

6.3.1. Numérica - Numérica

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



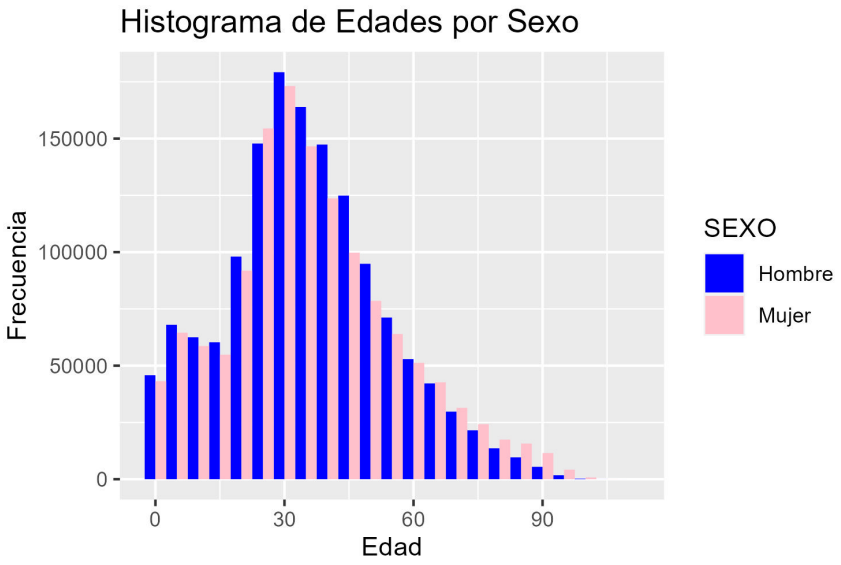
Efectivamente, como era de suponer, obtenemos que las dos variables están correlacionadas totalmente, ya que EDAD= 2021 - ANONAC.

```
       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 ambas variables es lineal.

6.3.2. 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.

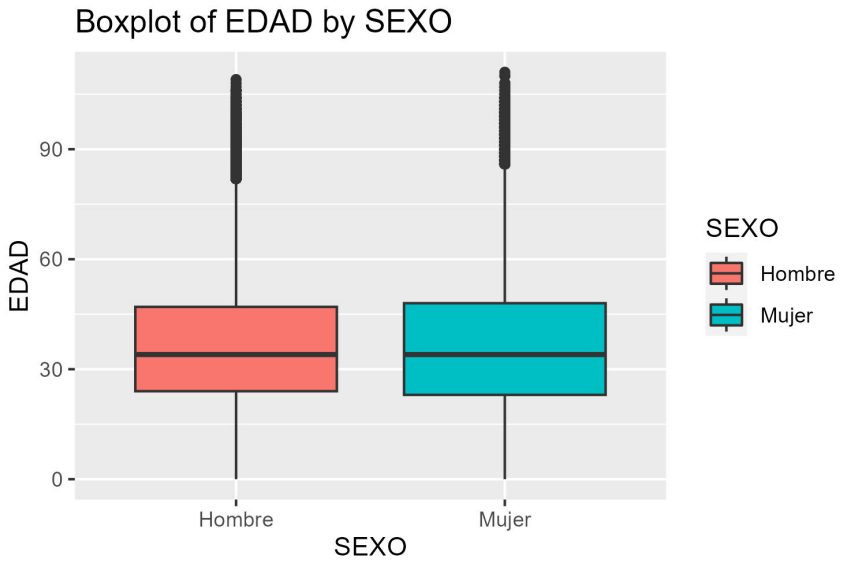


En esta gráfica, se observa que la gente cambia más de residencia alrededor de los 30 años, lo cual tiene sentido, ya que es la etapa de la vida en la que muchas personas deciden independizarse o formar una nueva familia.

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 que tras el nacimiento de los hijos o en muchas ocasiones del segundo hijo de la pareja, 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
```

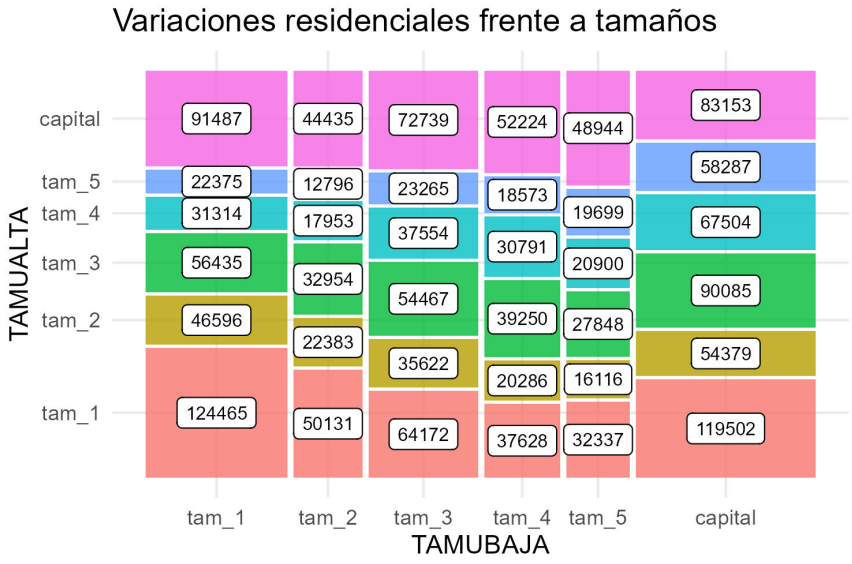


Por tanto, rechazamos la hipótesis nula de que las medias de edad de los hombres y las mujeres es la misma.

6.3.3. 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:

Código en el mosaico | Descripción |
-:- | -:- |
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 |

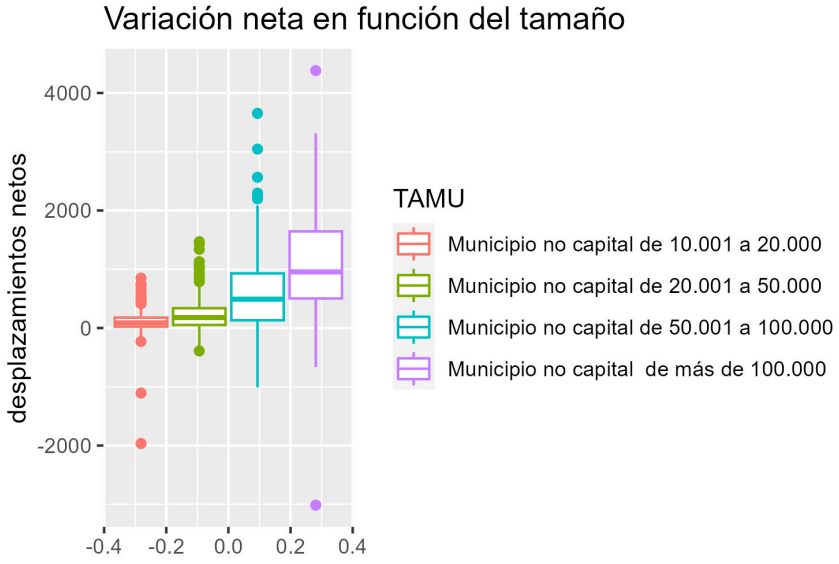


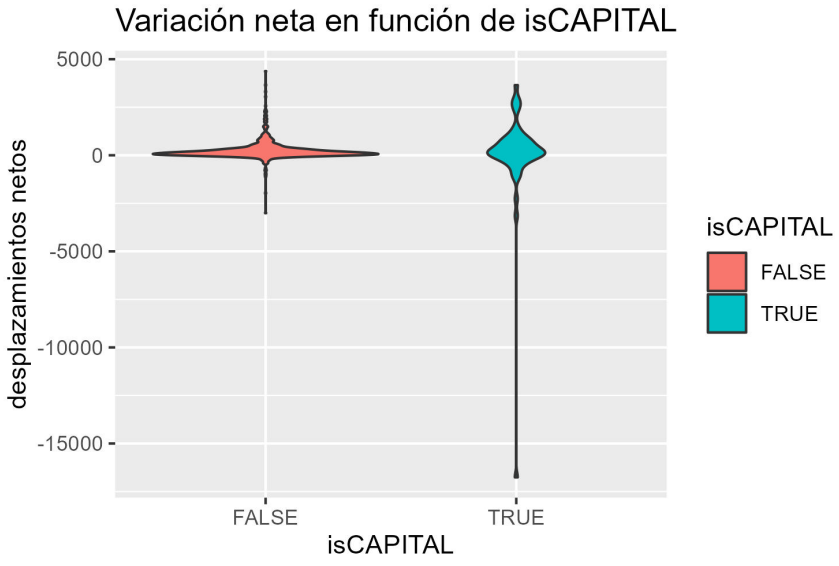
Para complementar este análisis, transformamos nuestros datos a fin de obtener un `data.frame` con la siguiente estructura:

- **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.

6.4. 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 \parencite(ggmap) 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 el documento `ProyectoAED2023.Rmd`.

6.4.1. Características

Se usa el test Chi-cuadrado. Este test supone una hipótesis de partida Ho (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.

6.5. Análisis de outliers y datos faltantes

Respecto a los valores NA, hemos decidido no imputar ningún valor faltante, ya que consideramos que nuestra muestra es suficientemente grande como para poder trabajar sin dichos valores y es preferible no introducir datos que modifiquen las características de nuestro conjunto de datos.

Respecto a los outliers, únicamente hemos encontrado en la variable EDAD aunque no se pueden considerar valores imposibles, puesto que una persona puede vivir 110 años.

```
  Media Mediana IQR Q1 Q3 ValorMinBoxplot
1  35.7      34  23 24 47           -10.5
  ValorMaxBoxplot
1            81.5

  Media Mediana IQR Q1 Q3    intervalo
1  35.7      34  23 24 47 -10.5 - 81.5

      n
1 54238

Adding missing grouping variables: `SEXO`

# A tibble: 2 x 8
  SEXO   Media Mediana   IQR    Q1    Q3
  <fct>  <dbl>   <dbl> <dbl> <dbl> <dbl>
1 Hombre  35.4      34    23    24    47
2 Mujer   36.1      34    25    23    48
# i 2 more variables: ValorMinBoxplot <dbl>,
#   ValorMaxBoxplot <dbl>

# A tibble: 2 x 7
  SEXO   Media Mediana   IQR    Q1    Q3 intervalo
  <fct>  <dbl>   <dbl> <dbl> <dbl> <dbl> <chr>
1 Hombre  35.4      34    23    24    47 -10.5 - ~
2 Mujer   36.1      34    25    23    48 -14.5 - ~
```

```
# A tibble: 1 x 2
# Groups:   SEXO [1]
  SEXO      n
  <fct> <int>
1 Mujer 22619

# A tibble: 1 x 2
# Groups:   SEXO [1]
  SEXO       n
  <fct>  <int>
1 Hombre 19145

  Sigma
1  93.5

     n
1 5058

Adding missing grouping variables: ‘SEXO‘

# A tibble: 2 x 2
  SEXO   Sigma
  <fct>  <dbl>
1 Hombre  91.3
2 Mujer   95.8

# A tibble: 1 x 2
# Groups:   SEXO [1]
  SEXO       n
  <fct>  <int>
1 Hombre  2725

# A tibble: 1 x 2
# Groups:   SEXO [1]
  SEXO      n
  <fct> <int>
1 Mujer  1826

[1] 53.4
```

La regla del identificador de Hampel es el único que no considera que la distribución sea gaussiana. Sin embargo, etiqueta como outliers valores que realmente no lo son. 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 outliers.

Supplementary Materials: No hay información de apoyo disponible.

Author Contributions: S.O. y J.H. hicieron la búsqueda y selección del datasetM 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