*Article*

# Análisis exploratorio de un conjunto de datos de indicadores urbanos

**Rebeca Company[1,*], Alejandro Dionis[1,*], Gabriel Ivars[1,*], Julio Garcia[1,*]**

[1] Universitat de València - ETSE-UV, Avinguda de l'Universitat, 46100 Burjassot, Valencia;
[*] Correspondence: etse@uv.es; Tel.: +34-963-54-3211.

**Abstract:** El presente artículo es un proyecto de la Asignatura Análisis Exploratorio de Datos, en el que se realiza un análisis exploratorio del conjunto de datos de Indicadores Urbanos, publicado por el Instituto Nacional de Estadística. A lo largo del mismo se plantean y responden preguntas de interés sobre aspectos demográficos, económicos y sociales de España en los últimos 15 años.

**Keywords:** Indicadores Urbanos; Tasa ; Población

## 1. Introducción

Hablar de los 3 datasets y de donde se han obtenido. BLABLABLA

## 2. Objetivos

## 3. Análisis exploratorio de los datos

### *3.1. Importación de los datos*

Antes de importar, primero se verificó que la codificación de los tres datasets fuera la misma, confirmando que todos estaban en formato UTF-8. Una vez comprobado, a la hora de importar los datasets, fue necesario establecer el delimitador de campos con el punto y coma (;) y adaptar el formato original al de R, ya que en los datasets el decimal_mark es la coma (,) y el grouping_mark es el punto (.). Las variables a estudiar, las cuales son coincidentes en los tres datasets, son:

- *Total Nacional* : Variable redundante con un único valor "Total Nacional", que toma verdadero valor cuando el atributo *Municipios* es vacío.
- *Municipios* : Municipio del que se han obtenido los datos.
- *Indicadores* : Tipo de estadístico demográfico, económico o social que se está calculando en cada observación. Se comentarán mas a fondo en los hallazgos obtenidos, ya que se trata de un elevado número de indicadores.
- *Sexo* : Sexo del cual se está obteniendo el indicador estadístico (hombre o mujer).
- *Periodo* : Año al que hace referencia el indicador estadístico.
- *Total* : Valor numérico asociado al indicador estadístico, que puede ser un valor absoluto, una tasa o un porcentaje, dependiendo del tipo de estadístico.

Otras consideraciones que se tuvieron en cuenta fué que la variable *Total* del dataset social era de tipo character y se transformó a tipo numeric para poder realizar operaciones. Esto se debe a que esta columna contenía valores "..", que serán considerados como faltantes en nuestro análisis. Al importar estos valores, R determinó que la clase de esta variable era de tipo character.
Además, se transformó la variable *Periodo* a tipo Date y la variable *Sexo* a tipo factor.
Posteriormente se concatenaron los tres datasets ya que tienen formato tidy.
Finalmente, el dataset se separó en varias partes:

**Citation:** Company, R.; Dionis, A.; Ivars, G.; Garcia, J. Análisis exploratorio de un conjunto de datos de indicadores urbanos. *Journal Not Specified* **2023**, *1*, 0. https://doi.org/

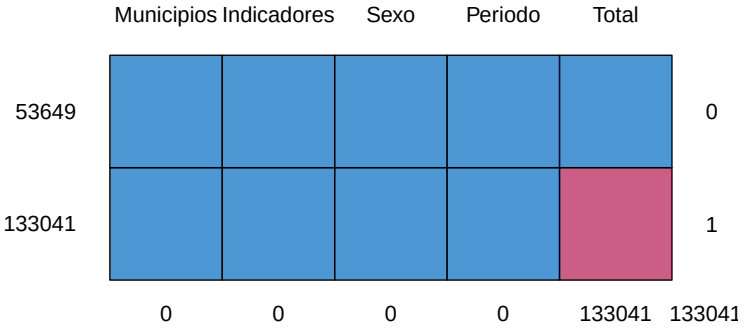
Received:
Revised:
Accepted:
Published:

- Datos totales nacionales con *Sexo* = "Total".
- Datos totales nacionales por sexo.
- Datos por municipios con *Sexo* = "Total".
- Datos por municipios y sexo.

Esto se hizo porque no tenía sentido estudiar los datos de forma conjunta, ya que muchos de los indicadores tienen distintas escalas. No obstante, también se ha guardado el dataset completo ya que es necesario para el estudio de los datos faltantes.

### *3.2. Análisis de datos faltantes*

En esta sección previa al análisis univariante y bivariante se va a realizar un análisis de la estructura de los faltantes y del dataset en general. Para asegurar que nuestro conjunto de datos sea coherente, evaluaremos el porcentaje de valores faltantes (NA) presentes en cada variable:

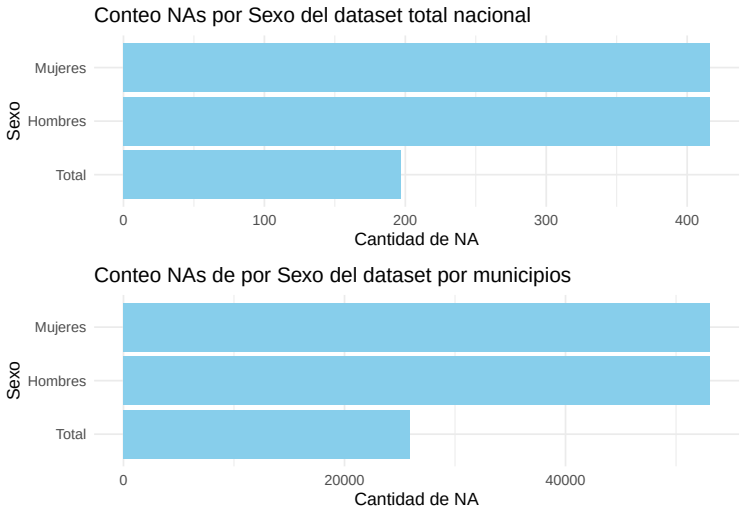



```
##        Municipios Indicadores Sexo Periodo  Total
## 53649           1           1    1       1      1      0
## 133041          1           1    1       1      0      1
##                 0           0    0       0 133041 133041
```

| | x |
|---|---|
| Municipios | 0.00000 |
| Indicadores | 0.00000 |
| Sexo | 0.00000 |
| Periodo | 0.00000 |
| Total | 71.26306 |

Del dataset completo, es decir, sin discriminar por total nacional ni sexo, únicamente se observan valores faltantes en la variable *Total*, siendo este un valor muy elevado (71.26%). Es importante señalar que durante la importación de los datasets, se eliminaron los datos faltantes en la variable Municipios que correspondían a los valores asociados a "Total Nacional". Por esta razón, se eliminó la columna Total Nacional y se rellenaron los valores NA de *Municipios* con el valor "Total Nacional".
También se observó que al modificar el tipo de la variable *Total* se apreció un incremento del número de valores faltantes. Esto se debió a las 1461 observaciones del dataframe social con un valor de ".." que al transformarlos a numéricos, R los asoció a valores NA.
Además, no se encontraron observaciones duplicadas a lo largo del dataset completo.
A continuación se van a estudiar estos valores faltantes en función del resto variables no numéricas.

Por sexo



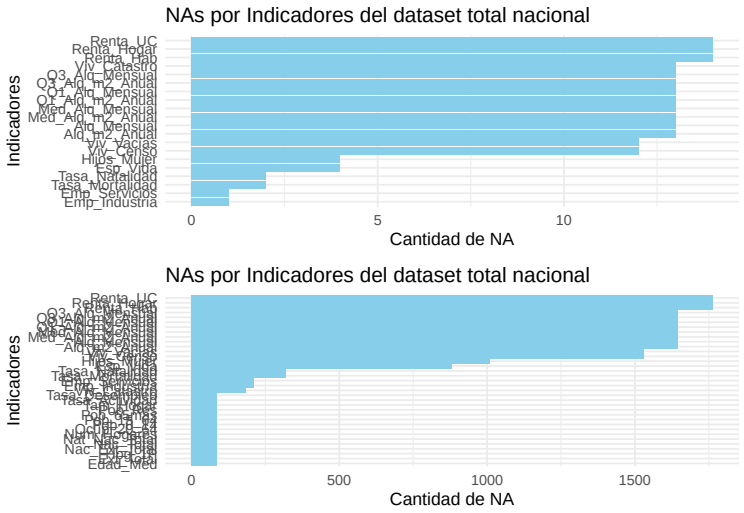

```
## # A tibble: 3 x 4
##   Sexo        n  perc perc_total
##   <fct>   <int> <dbl>      <dbl>
## 1 Hombres   416  84.9       28.3
## 2 Mujeres   416  84.9       28.3
## 3 Total     197  40.2       13.4
```

```
## # A tibble: 3 x 4
##   Sexo        n  perc perc_total
##   <fct>   <int> <dbl>      <dbl>
## 1 Hombres 53046  85.9       28.6
## 2 Mujeres 53046  85.9       28.6
## 3 Total   25920  42.0       14.0
```

Tanto para hombres como para mujeres, aproximadamente el 85% de sus observaciones tienen valores faltantes en la columna *Total*, lo que representa mas o menos el 28% del total de observaciones en ambos conjuntos de datos. Para el nivel Total (variable *Sexo*), en torno al 40% de sus observaciones tienen valores faltantes, representando el 14% aproximandamente del total de observaciones en ambos dataframes. Por tanto, debido al elevado porcentaje de valores faltantes en las observaciones desglosadas por sexo, decidimos no incluir en el análisis observaciones de hombres y mujeres por separado y optamos por trabajar únicamente con el total combinado de ambos sexos. Por tanto, a partir de ahora se va a realizar el análisis sobre estos dos datasets:

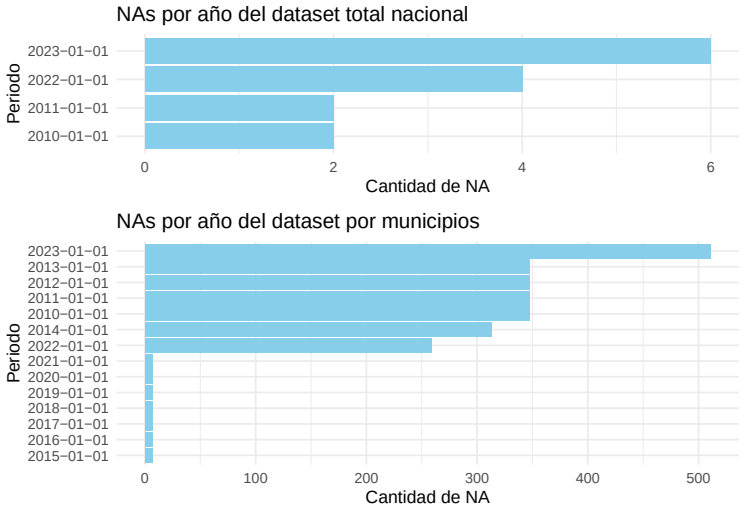
- Datos totales nacionales con *Sexo* = "Total".
- Datos por municipios con *Sexo* = "Total".

Por indicadores




Como se observa en la gráfica superior, la distribución de los NAs por cada indicador no es uniforme en los dos datasets. Por tanto, como criterio, se van a prescindir de los que superen el 50% de valores faltantes relativo al total de observaciones para cada indicador.

Por periodo



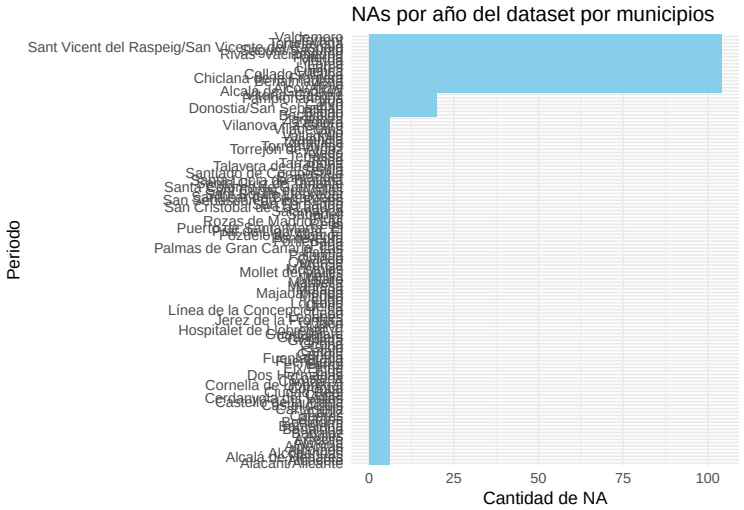

```
## # A tibble: 5 x 5
##   Periodo    n_total  n_na  perc perc_total
##   <date>       <int> <dbl> <dbl>      <dbl>
## 1 2023-01-01      21     6 28.6       2.04
## 2 2022-01-01      21     4 19.0       1.36
## 3 2010-01-01      21     2  9.52      0.680
## 4 2011-01-01      21     2  9.52      0.680
## 5 2012-01-01      21     0  0         0
```

```
## # A tibble: 8 x 5
##   Periodo     n_na n_total   perc perc_total
##   <date>     <int>   <int>  <dbl>      <dbl>
## 1 2023-01-01   511    2520 20.3       1.45
## 2 2010-01-01   347    2520 13.8       0.984
## 3 2011-01-01   347    2520 13.8       0.984
## 4 2012-01-01   347    2520 13.8       0.984
```

```
## 5 2013-01-01   347    2520 13.8       0.984
## 6 2014-01-01   313    2520 12.4       0.887
## 7 2022-01-01   259    2520 10.3       0.734
## 8 2015-01-01     7    2520  0.278     0.0198
```

El análisis de los datos faltantes revela patrones interesantes. En ambos conjuntos de datos, el año 2023 presenta la mayor proporción de valores faltantes. Sin embargo, en el dataset municipal, se observa un período de estabilidad con un bajo porcentaje de datos faltantes entre 2010 y 2013. A pesar de estas variaciones, el porcentaje total de datos faltantes es relativamente bajo, inferior al 2%. Por lo tanto, se procederá a imputar estos valores para completar el conjunto de datos.

### 3.2.1. Por municipios




```
## # A tibble: 126 x 5
##    Municipios              n_na n_total  perc perc_total
##    <fct>                  <int>   <int> <dbl>      <dbl>
##  1 Alcalá de Guadaíra       104     280  37.1      0.295
##  2 Alcoi/Alcoy              104     280  37.1      0.295
##  3 Ávila                    104     280  37.1      0.295
##  4 Benalmádena              104     280  37.1      0.295
##  5 Chiclana de la Frontera  104     280  37.1      0.295
##  6 Collado Villalba         104     280  37.1      0.295
##  7 Cuenca                   104     280  37.1      0.295
##  8 Linares                  104     280  37.1      0.295
##  9 Lorca                    104     280  37.1      0.295
## 10 Mérida                   104     280  37.1      0.295
## # i 116 more rows
```

Finalmente se observa una mejora de valores faltantes al prescindir de la variable Sexo y de los indicadores con mayor proporción de NAs:

```
## Indicadores     Periodo       Total         Ind
##    0.000000    0.000000    4.761905    0.000000

##  Municipios Indicadores     Periodo       Total         Ind
##    0.000000    0.000000    0.000000    7.142857    0.000000
```

Estos valores van a ser imputados mediante una medida de tendencia central, dependiendo del indicador al que esté asociada cada observación. Destacar también que se estudió la existencia de observaciones duplicadas, obteniendo ausencia de estas.

Imputación

**4. USAR MICE ?**

```
## # A tibble: 1 x 4
##   statistic    df p.value missing.patterns
##       <dbl> <dbl>   <dbl>            <int>
## 1      325.     4       0                2
```

Dado que el valor p es 0, esto sugiere que los datos faltantes no son MCAR. Esto significa que los datos faltantes podrían ser MAR (Missing at Random) o MNAR (Missing Not at Random).
Para asegurar la consistencia de los datos, se realizó una imputación de los valores faltantes en la columna *Total*. Para llevar a cabo la imputación, se realizó por separado en dos datasets: uno para los valores separados por municipios, y otro para el total nacional. Después, para cada dataset se calcularon tanto la media como la mediana de *Total* para cada grupo de Indicadores. Luego, se compararon estas dos medidas y se observó que, en ciertos indicadores, la diferencia entre la media y la mediana era significativa. Debido a estas discrepancias, se decidió utilizar la mediana para la imputación, ya que es menos sensible a los valores atípicos. Finalmente, se imputaron los valores faltantes en la columna *Total* con la mediana correspondiente a cada grupo de Indicadores.

### *4.1. Análisis univariante*

#### 4.1.1. Características generales

En esta sección se lleva a cabo un análisis preeliminar de los dos datasets que van a ser estudiados, prestando especial atención a las posibles anomalías, como datos inconsistentes u outliers.
Se estudian paralelamente los dos dataframes: uno de ellos contiene los valores de los indicadores para cada municipio (37044 observaciones), y el otro los valores del total nacional (294 observaciones). Ambos sets de datos contienen las siguientes variables:

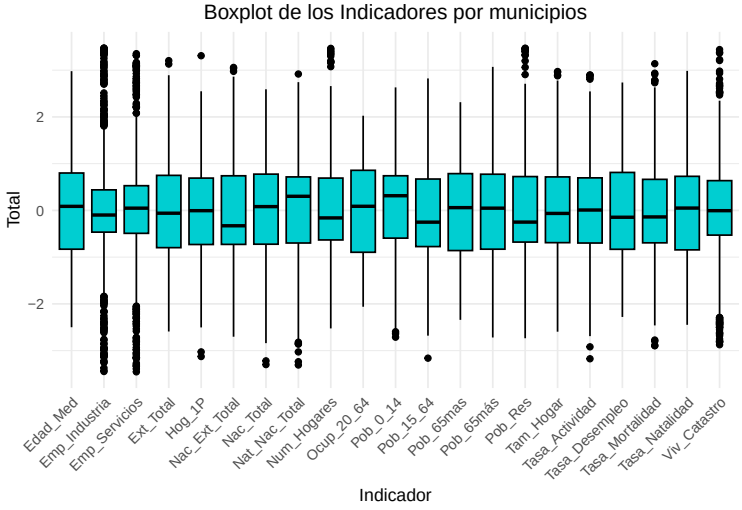
| Variable | Tipo | Niveles | Valores |
|---|---|---|---|
| Municipios | Texto | 126 | Albacete, Alcalá de Guadaíra, . . . , Zaragoza |
| Indicadores | Texto | 42 | Pob_Res, . . . , Pob_65más |
| Periodo | Fecha | 14 | 2010-01-01 - 2023-01-01 |
| Total | Numérica | N/A | NA - NA |

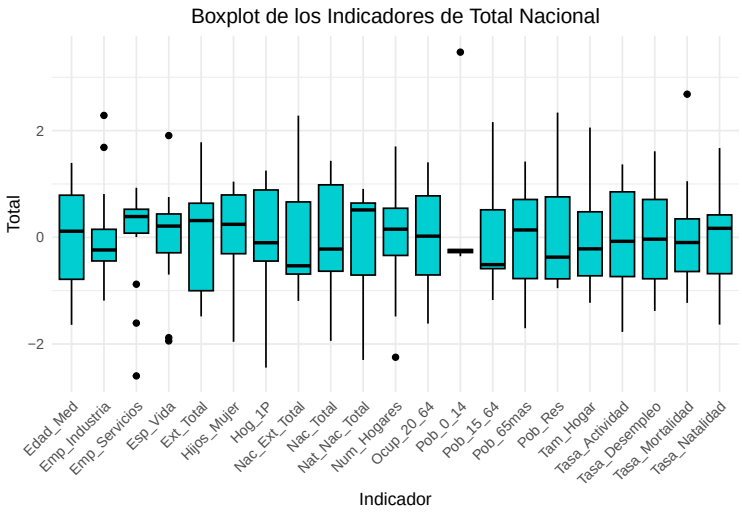
- Municipios: Municipio del que se han obtenido los datos.
- Indicadores: Estadístico demográfico, económico o social que se está calculando.
- Periodo: Año al que hace referencia el indicador estadístico.
- Total:Valor numérico asociado al indicador estadístico, que puede ser un valor absoluto, una tasa o un porcentaje.

Sin embargo, para nuestro análisis es relevante considerar cada indicador como una variable individual, en lugar de mantenerlos agrupados en la columna *Indicadores*, optamos por transformar los datasets a formato wide, de manera que quede cada indicador como una variable nueva.
En el caso de las variables *Municipio* y *Periodo*, son consideradas como variables no numéricas y se han revisado sus valores únicos y sus frecuencias absolutas y relativas, para asegurar que no hay ninguna anomalía.
Respecto a las variables numéricas, se revisan sus principales estadísticos descriptivos, para detectar posibles valores negativos, tasas o proporciones mayores que 100, etc... De nuevo no se detectan inconsistencias.

#### 4.1.2. Exploración visual

Para buscar valores anómalos se decide graficar los indicadores en boxplots.




Se considera relevante estudiar individualmente aquellos indicadores que parecen susceptibles de presentar datos anómalos:

- De los datos por municipios, ninguno de los indicadores parece presentar valores anómalos.
- De los datos del Total Nacional: *Esperanza de Vida*, *Número de Hogares*,, *Proporción de Población entre 0 y 14 años*, *Proporción de Empleos en Industria*, *Proporción de Empleo en Servicios y Tasa de Mortalidad*.

En el caso del dataset del Total Nacional:

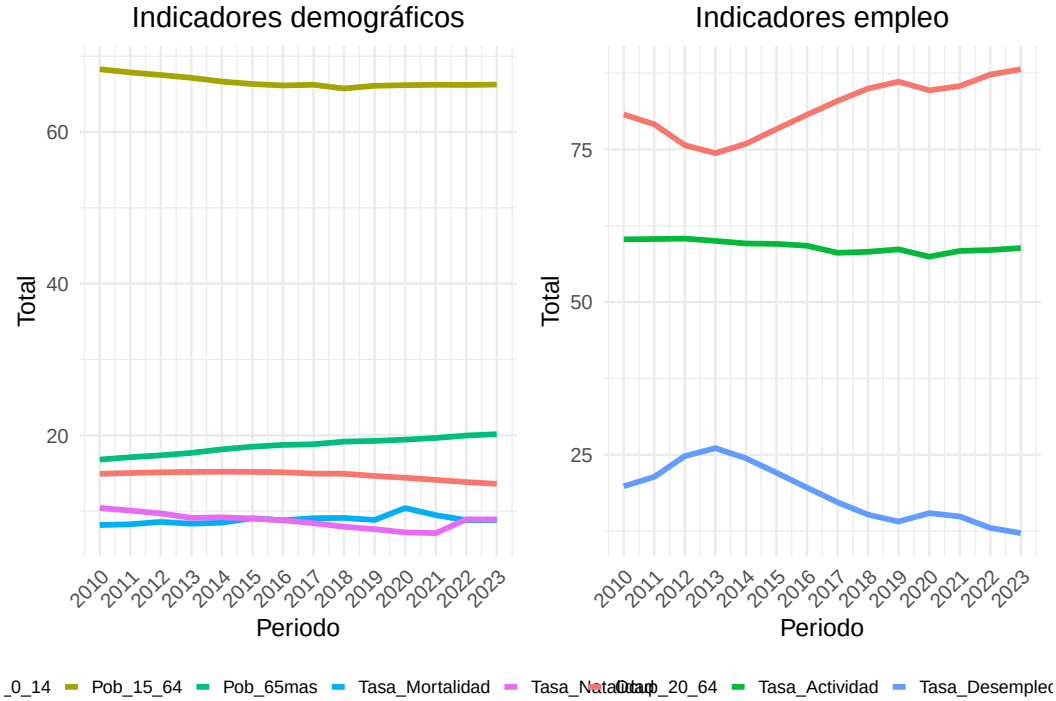
- En el caso de la variable *Esperanza de Vida* vemos que hay valores que superan los "límites" superior e inferior, ya que siguen siendo muy similares al resto, pero la desviación típica de esta variable es muy pequeña, 0.3298826. Por lo tanto tampoco se consideran valores anómalos.
- En las variables *Proporción de Empleo en Servicios* y *Proporción de Empleo en Industria*, observamos una situación similar a la anterior (variables con desviación típica pequeña y rango reducido). Además, los valores más "extremos" corresponden a los mismos años, pero se comportan de manera inversa entre sí. De nuevo, estos valores no se consideran anómalos.

| Variable | SD | Mínimo | Máximo |
|---|---|---|---|
| Emp_Industria | 0.4004373 | 11.94 | 13.33 |
| Emp_Servicios | 1.2218759 | 73.36 | 77.67 |

- En el caso de la variable *Proporción de población de 0-14 años*, el dato 61.43 sí se considera un dato anómalo, ya que es un valor muy alto en comparación al resto (el resto de valores de la variable oscilan entre 13.6, 15.19%). Además, para cada observación, la suma de las variables *Prop_0_14*, *Prop_15_64* y *Prop_65* da un resultado muy cercano al 100%, pero para en el caso del dato anómalo la suma resulta en 127.66. Se decide corregir este valor para que la suma de las tres proporciones sea 100.
- Finalmente revisamos el valor máximo de la variable *Tasa de Mortalidad*. Aunque es superior a los demás, la diferencia no es excesiva. Este valor corresponde al año 2020, por lo que el aumento en la tasa de mortalidad podría estar relacionado con la pandemia de COVID-19. A pesar de que este dato podría tener un impacto significativo, no se considera un valor anómalo.

Como parte del análisis univariante, también se lleva a cabo un estudio sobre las distribuciones de los diferentes indicadores a lo largo del tiempo. Para este análisis, se utiliza exclusivamente el dataset de Total Nacional, y se destacan algunos puntos interesantes:

- Indicadores demográficos: Los datos muestran claramente una tendencia al envejecimiento de la población. A lo largo de los años, la proporción de población de 0 a 14 años disminuye de forma notable, mientras que aumentan las proporciones de población de 15 a 64 años y de más de 64 años. Esto se ve, reforzado por un descenso de la Tasa de Natalidad (con un pequeño repunte en los últimos años.
- Indicadores económicos: En cuanto a los indicadores económicos, se destacan los relacionados con el empleo. Como era de esperar, la Tasa de Desempleo y la Tasa de Ocupados (de 20 a 64 años) siguen tendencias opuestas con el paso de los años. Durante el periodo analizado, se aprecia una fase de recesión laboral hasta 2013, y tras el final de la crisis de la burbuja inmobiliaria (2008-2014), mejoran las tasas de desempleo y de ocupación, aunque sufren un pequeño descenso en 2019.



#### 4.1.3. Estadísticos descriptivos

Una vez eliminado e imputado el dato anómalo encontrado, se vuelven a calcular los estadisticos descriptivos de las variables numéricas, es decir, de todos los indicadores de ambos datasets. Las tablas con los estadísticos descriptivos del dataset por municipios y del dataset total se muestran en las tablas **??** y **??** del Anexo 1.

### *4.2. Análisis bivariante*

En el Anexo 1, en la Tabla **??**

## 5. Anexo 1

**Table 4.** Estadísticos Dataset por Municipios

| Indicador | Min | Q1 | Media | Mediana | Q3 | Max | SD |
|---|---|---|---|---|---|---|---|
| Pob_Res | 51128.00 | 76136.00 | 190424.30 | 102717.50 | 196904.75 | 3340176.00 | 329107.98 |
| Pob_0_14 | 9.73 | 13.37 | 15.09 | 15.12 | 16.55 | 23.25 | 2.34 |
| Pob_15_64 | 60.52 | 65.62 | 67.20 | 67.22 | 68.54 | 78.13 | 2.56 |
| Pob_65mas | 5.80 | 15.04 | 17.76 | 17.61 | 20.48 | 29.57 | 4.19 |
| Edad_Med | 33.38 | 40.13 | 42.50 | 42.47 | 44.59 | 51.74 | 3.26 |
| Nac_Total | 46.78 | 85.44 | 89.10 | 90.57 | 94.37 | 98.93 | 7.19 |
| Nat_Nac_Total | 44.47 | 81.39 | 85.48 | 86.75 | 90.77 | 98.05 | 7.94 |
| Nac_Ext_Total | 1.95 | 9.03 | 14.44 | 13.12 | 18.61 | 55.53 | 7.98 |
| Ext_Total | 1.07 | 5.61 | 10.88 | 9.36 | 14.54 | 53.22 | 7.21 |
| Tasa_Natalidad | 4.54 | 7.27 | 8.59 | 8.33 | 9.76 | 18.43 | 1.94 |
| Tasa_Mortalidad | 2.36 | 6.90 | 8.24 | 8.16 | 9.53 | 14.65 | 2.12 |
| Tasa_Desempleo | 5.18 | 13.51 | 19.68 | 18.68 | 24.31 | 46.67 | 7.79 |
| Ocup_20_64 | 53.33 | 75.83 | 80.31 | 81.40 | 86.50 | 94.83 | 7.85 |
| Tasa_Actividad | 48.02 | 56.36 | 58.97 | 58.97 | 61.29 | 72.91 | 4.20 |
| Emp_Servicios | 49.91 | 77.47 | 82.13 | 83.15 | 87.51 | 95.70 | 7.33 |
| Emp_Industria | 0.96 | 4.65 | 9.37 | 7.49 | 12.99 | 42.03 | 6.58 |
| Num_Hogares | 18822.00 | 28682.00 | 73123.03 | 40039.00 | 73069.50 | 1332826.00 | 129658.07 |
| Tam_Hogar | 2.15 | 2.47 | 2.63 | 2.63 | 2.75 | 3.55 | 0.23 |
| Hog_1P | 12.33 | 22.95 | 26.18 | 26.19 | 29.75 | 41.67 | 4.99 |
| Viv_Catastro | 20225.00 | 33764.50 | 88915.45 | 51568.00 | 89973.00 | 1510442.00 | 149005.52 |
| Pob_65más | 5.80 | 15.02 | 17.71 | 17.56 | 20.46 | 29.57 | 4.17 |

**Table 5.** Estadísticos Dataset Total Nacional

| Indicador | Min | Q1 | Media | Mediana | Q3 | Max | SD |
|---|---|---|---|---|---|---|---|
| Pob_Res | 46440099.00 | 46527182.25 | 46916416.14 | 46730464.00 | 47295382.25 | 48085361.00 | 499893.26 |
| Pob_0_14 | 13.60 | 14.46 | 14.73 | 14.94 | 15.13 | 15.19 | 0.53 |
| Pob_15_64 | 65.74 | 66.19 | 66.64 | 66.25 | 67.03 | 68.28 | 0.76 |
| Pob_65mas | 16.80 | 17.80 | 18.63 | 18.77 | 19.39 | 20.15 | 1.07 |
| Edad_Med | 39.86 | 41.39 | 42.80 | 43.01 | 44.21 | 45.30 | 1.79 |
| Nac_Total | 87.34 | 88.53 | 89.11 | 88.90 | 90.00 | 90.41 | 0.91 |
| Nat_Nac_Total | 82.94 | 84.98 | 85.89 | 86.54 | 86.71 | 87.05 | 1.28 |
| Nac_Ext_Total | 12.68 | 13.32 | 14.18 | 13.51 | 15.02 | 17.06 | 1.26 |
| Ext_Total | 9.59 | 10.04 | 10.98 | 11.28 | 11.59 | 12.66 | 0.94 |
| Tasa_Natalidad | 7.10 | 8.06 | 8.74 | 8.91 | 9.16 | 10.42 | 1.00 |
| Tasa_Mortalidad | 8.17 | 8.51 | 8.87 | 8.81 | 9.07 | 10.40 | 0.57 |
| Esp_Vida | 82.26 | 82.81 | 82.90 | 82.97 | 83.04 | 83.53 | 0.33 |
| Hijos_Mujer | 1.18 | 1.26 | 1.28 | 1.29 | 1.32 | 1.33 | 0.05 |
| Tasa_Desempleo | 12.18 | 14.98 | 18.59 | 18.43 | 21.89 | 26.09 | 4.65 |

| Indicador | Min | Q1 | Media | Mediana | Q3 | Max | SD |
|---|---|---|---|---|---|---|---|
| Ocup_20_64 | 74.38 | 78.53 | 81.74 | 81.84 | 85.27 | 88.13 | 4.55 |
| Tasa_Actividad | 57.44 | 58.42 | 59.11 | 59.04 | 59.92 | 60.40 | 0.94 |
| Emp_Servicios | 73.36 | 76.63 | 76.53 | 77.01 | 77.18 | 77.67 | 1.22 |
| Emp_Industria | 11.94 | 12.24 | 12.41 | 12.32 | 12.47 | 13.33 | 0.40 |
| Num_Hogares | 17520072.00 | 18159784.00 | 18273694.29 | 18324650.00 | 18456125.00 | 18844567.00 | 335153.28 |
| Tam_Hogar | 2.49 | 2.51 | 2.54 | 2.53 | 2.56 | 2.62 | 0.04 |
| Hog_1P | 18.99 | 24.36 | 25.56 | 25.28 | 27.94 | 28.92 | 2.69 |

**Author Contributions:** Todos los autores contribuyeron de manera equitativa a este trabajo.

**Data Availability Statement:** Los conjuntos de datos analizados en el presente artículo están disponibles en la sección "Indicadores Urbanos. Últimos datos." del repositorio público de conjuntos de datos abiertos del Instituto Nacional de Estadística INEbase https://www.ine.es/dyngs/INEbase/listaoperaciones.htm.

**Conflicts of Interest:** Los autores declaran no tener ningún conflicto de interés

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| MDPI | Multidisciplinary Digital Publishing Institute |
| INE | Instituto Nacional de Estadística |