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, Análisis Exploratorio

1. Introducción

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

2. Objetivos

El objetivo principal de este trabajo es llevar a cabo un análisis exploratorio de los conjuntos de datos seleccionados, con el fin de identificar posibles anomalías y comprender los aspectos más relevantes de la información contenida. Este análisis servirá como base para futuros estudios o para la aplicación de métodos más avanzados.
Asimismo, se busca responder a las siguientes preguntas clave a partir de los datos analizados:

- ¿Reflejan los datos algunos de los eventos más significativos en España, como la crisis económica de 2008-2014 o la pandemia de COVID-19 en 2019?
- ¿Que podemos deducir sobre la situación demográfica de España a partir de estos datos?

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 (35 niveles).
- *Sexo* : Sexo del cual se está obteniendo el indicador estadístico (hombre o mujer).

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:

- *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 "factor".
Posteriormente se concatenaron los tres datasets ya que estaban en formato "tidy". y finalmente, el dataset se separó en cuatro partes:

- 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 realizó un análisis de la estructura de los faltantes y del dataset en general. Para asegurar que nuestro conjunto de datos sea coherente, se evaluó el porcentaje de valores faltantes (*NA*) presentes en cada variable:

Table 1. Porcentajes de NAs

	Porcentaje
Municipios	0.00
Indicadores	0.00
Sexo	0.00
Periodo	0.00
Total	71.26

Del dataset completo, 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 *NAs* 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 transformar a "numeric", R asoció dichos valores a valores faltantes.
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*

Table 2. Porcentajes de NAs por sexo en el dataset Total Nacional

Sexo	n	Porcentaje	Porcentaje_total
Hombres	416	84.9	28.3
Mujeres	416	84.9	28.3
Total	197	40.2	13.4

Table 3. Porcentajes de NAs por sexo en el dataset por municipios

Sexo	n	Porcentaje	Porcentaje_total
Hombres	53046	85.91837	28.64
Mujeres	53046	85.91837	28.64
Total	25920	41.98251	13.99

Como se observa en la Tabla 2 y Tabla 3, 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,se decidió 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 esta decisión se continuó con el análisis de estos dos conjuntos de datos:

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

Por *Indicadores*, *Periodo* y *Municipios*

En ambos conjuntos de datos, la distribución de los valores faltantes variaba entre los diferentes niveles de las variables, alejándose considerablemente de una distribución uniforme. Como criterio, se eliminaron aquellos niveles de variables que tenían más del 50% de valores faltantes realtivo al número de observaciones de cada nivel.
Para el caso de la variable *Indicadores*, se eliminaron 14 y 15 indicadores del conjunto de datos nacional y del de municipios, respectivamente.
Se observaron patrones interesantes en la variable *Periodo*. En ambos conjuntos de datos, el año 2023 presentaba la mayor proporción de valores faltantes. Sin embargo, en el dataset municipal, se identificó entre 2010 y 2013 un bajo porcentaje de datos faltantes (13.77%), siendo igual en los tres años. No obstante, el porcentaje total de datos faltantes para todos los años es relativamente bajo, inferior al 2%. Por lo tanto, se decidió imputar estos valores para completar el conjunto de datos.
En el caso de la variable *Municipios*, se puede destacar que se identificaron 17 municipios con un porcentaje considerable de valores faltantes, específicamente del 37.14%. Dado que este porcentaje no supera el 50%, se decidió imputar estos valores en lugar de prescindir de dichos municipios.
Finalmente, se observó una mejora de valores faltantes al prescindir de la variable *Sexo* y de los indicadores con mayor proporción de *NAs*:

Table 4. Porcentajes de NAs conjunto Total Nacional

	x
Indicadores	0.00
Periodo	0.00
Total	4.76

Table 5. Porcentajes de NAs por municipios

	x
Municipios	0.00
Indicadores	0.00
Periodo	0.00
Total	7.14

Destacar también que se estudió la existencia de observaciones duplicadas, obteniendo ausencia de estas.

Imputación

Para la imputación, en primera instancia se optó por imputar ambos datasets mediante una medida de tendencia central, dependiendo del indicador al que estaba asociada cada observación. Tras la imputación, en el caso del conjunto por municipios, se observó que esta imputación sesgaba los datos y se decidió usar una técnica más sofisticada.
Como en el análisis parecían existir ciertos patrones de los faltantes con las variables del conjunto de datos, se realizó el test de Little's MCAR, rechazando así la hipótesis nula de que los datos faltantes eran completamente al azar (Missing Completely At Random). Esto confirmó nuestra hipótesis de que podrían depender de las variables observadas (Missing At Random). Tras estos resultados, se empleó la función *mice* con el método 'cart' (Classification and Regression Trees). Este método crea árboles de decisión para predecir los valores faltantes basándose en las relaciones entre las variables observadas. No obstante, tras esta imputación, para algunos municipios e indicadores, los valores faltantes estaban al principio de la serie temporal, sin datos previos disponibles. Por tanto, se imputaron de forma separada, utilizando extrapolación lineal.
Para el dataset por municipios 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.

3.3. Análisis Univariante

3.3.1. 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 (35280 observaciones), y el otro los valores del total nacional (294 observaciones). En la Tabla 6 podemos ver las variables y su tipo.

Table 6. Variables de ambos datasets

Variable	Tipo	Niveles	Valores
Municipios	Texto	126	Albacete, Alcalá de Guadaíra, . . . , Zaragoza

Variable	Tipo	Niveles	Valores
Indicadores	Texto	41	Pob_Res, . . . , Viv_Catastro
Periodo	Fecha	14	2010-01-01 - 2023-01-01
Total	Numérica	N/A	0.96 - 48085361

- 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.

3.3.2. Exploración visual

Para buscar valores anómalos se decide graficar los indicadores en boxplots (Figuras 1 y 2).

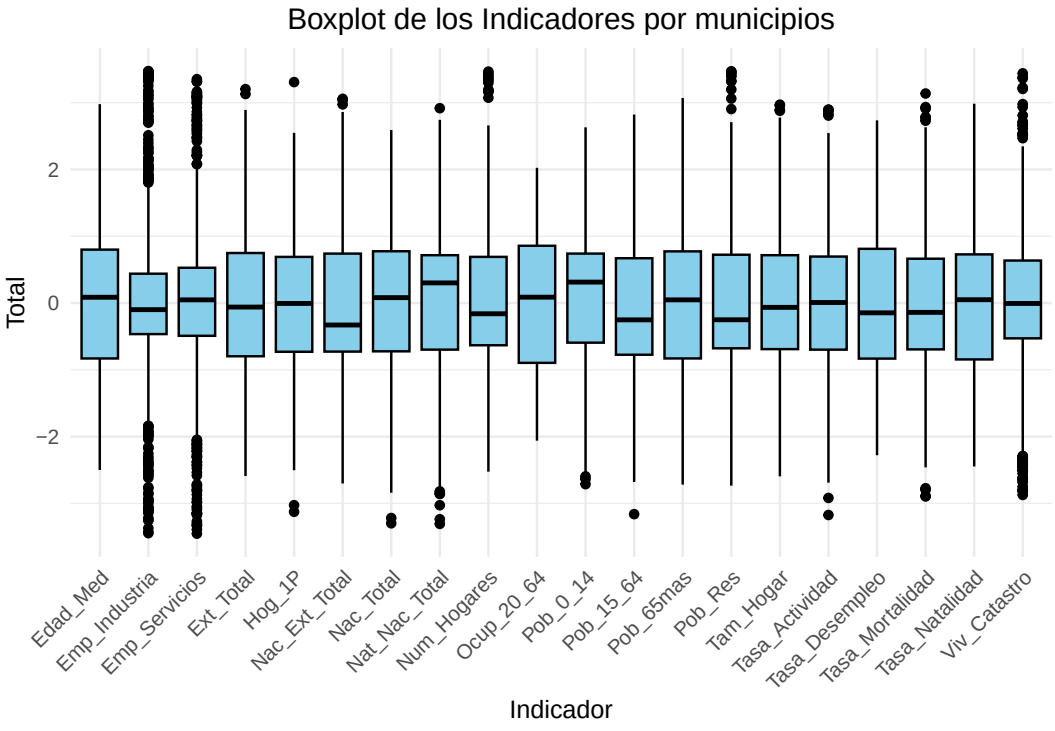

Figure 1. Boxplot múltiple del dataset por municipios

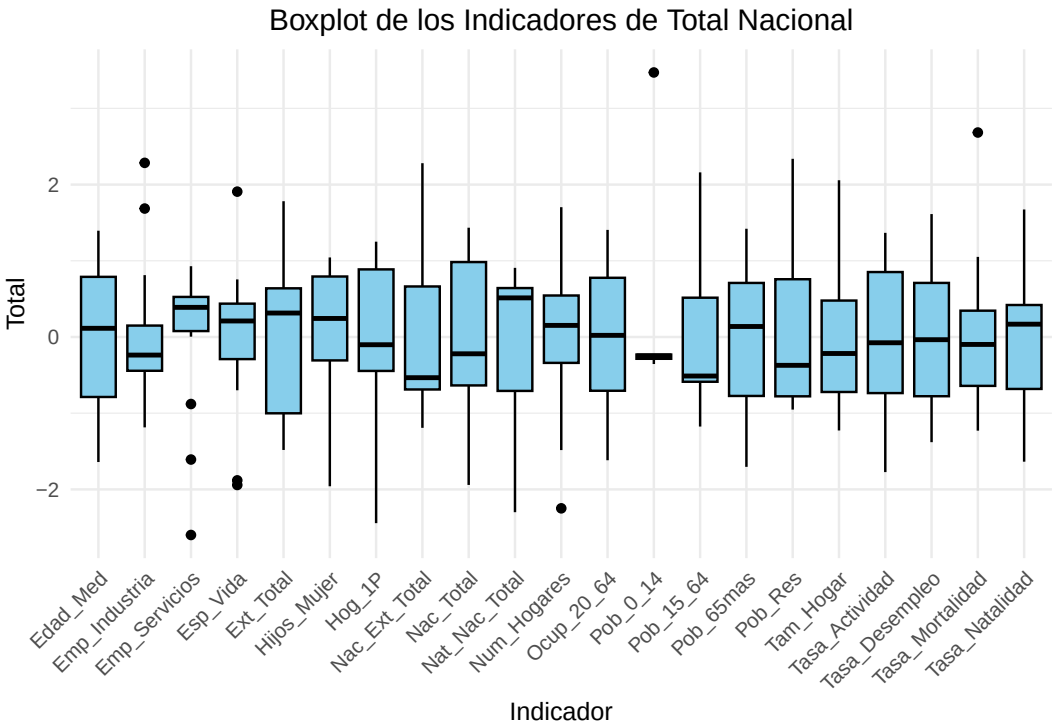

Figure 2. Boxplot múltiple del dataset total nacional

Se decide estudiar individualmente aquellos indicadores que visualmente 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 la variable *Esperanza de Vida* vemos que hay valores que superan los "límites" superior e inferior, a pesar de que siguen siendo muy similares al resto. Esto se debe a que esta variable presenta muy poca variabilidad (la desviación típica de esta variable es 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 poca variabilidad: 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.
- 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*. Este valor corresponde al año 2020, por lo que dicho 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 y se decide mantenerlo porque aporta información que puede ser de utilidad.

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, para obtener una visión general de todo el país y se destacan algunos puntos interesantes:

- Indicadores demográficos (Tabla 3 izquierda) : Los datos muestran claramente una tendencia al envejecimiento de la población. Esto se ve reflejado de forma direca en el aumento de la edad media, pero también en otros indicadores. 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 (Tabla 3 derecha) : 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.

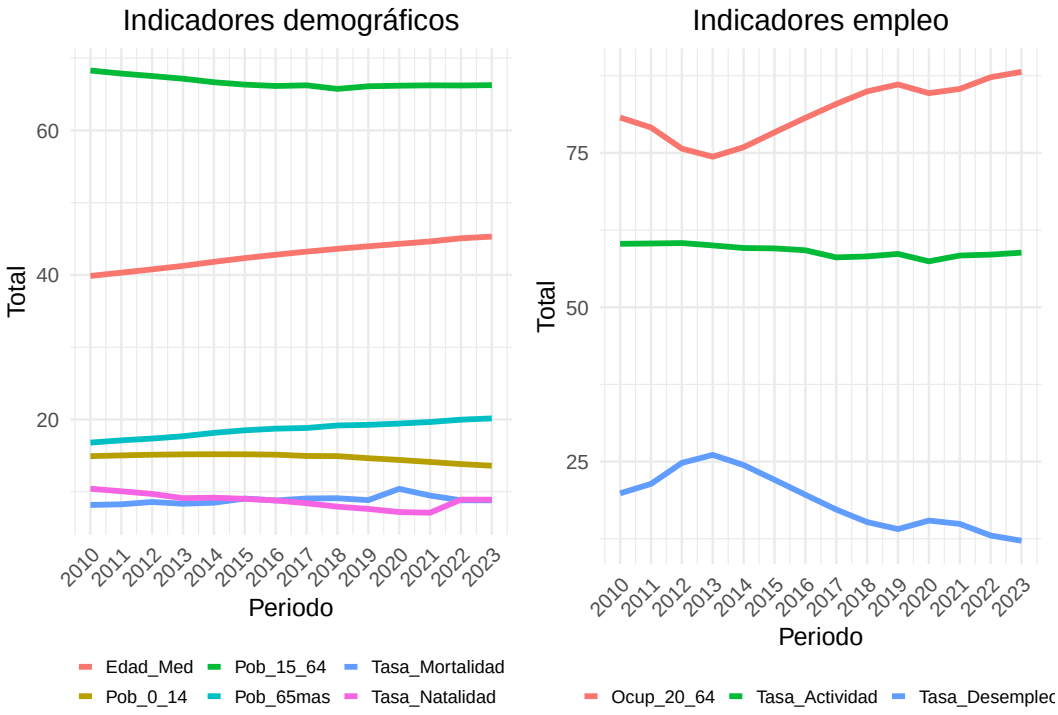

Figure 3. Estudio de los Indicadores respecto al PEriodo

3.3.3. 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 7 y 8 del Anexo 1.

3.4. Análisis bivariante

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

4. Anexo 1

Table 7. 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	15.02	17.71	17.56	20.46	29.57	4.17
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

Table 8. 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
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