FACULTAD DE INFORMÁTICA

PRÁCTICA

Tecnologías de Integración

Carlos Hermida

3º GCID – Curso 2022/2023

UNIVERSIDADE DA CORUÑA

Descripción del problema

Las viviendas de uso turístico (VUT) en Galicia han experimentado un crecimiento significativo en los últimos años. Se trata de propiedades gestionadas por particulares, alquiladas a turistas y viajeros de forma ocasional y por periodos de corta duración, como una alternativa a los establecimientos hoteleros tradicionales.

Debido a que cualquier propietario puede registrar su vivienda como VUT y anunciarla en sitios web como *Airbnb* o *Booking*, varios municipios gallegos se han llenado de este tipo de alojamientos.

En esta práctica, se integrarán diferentes fuentes de datos para averiguar cuántos turistas han visitado en mayo de 2023 aquellos municipios donde hay al menos una VUT por cada 100 habitantes.

municipio	numero_de_vt	habitantes	vt_cada_100_hab	turistas_nacional_mayo_2023	turistas_internacional_mayo_2023	turistas_total_mayo_2023
Sanxenxo	2624	17760	14.8	17740	2986	20726
Vilagarcía de Arousa	359	37677	1	11868	797	12665
O Grove	606	10809	5.6	9714	1065	10779
Boiro	231	18976	1.2	7724	460	8184
Ribeira	292	26897	1.1	6613	686	7299
Ribadeo	241	9811	2.5	6733	563	7296
Nigrán	290	18054	1.6	3821	1899	5720
Cangas	577	26832	2.2	4310	1136	5446
Viveiro	342	15231	2.2	4502	382	4884
Baiona	342	12349	2.8	3296	1532	4828

REST Web Service de Denodo mostrando el resultado final

Diseño de la integración

Fuentes de datos

1. **Alojamientos turísticos en Galicia**

- nombre del archivo: *reat_directorio-alojamientos_esp.csv*

- formato: CSV

- fecha: 01/07/2023

- url: https://aei.turismo.gal/es/rexistro-de-empresas-e-actividades-turisticas

Procedencia del archivo con los datos de alojamientos

Este archivo contiene los datos de todos los alojamientos turísticos (hoteles, pensiones, albergues, viviendas de uso turístico…) registrados en Galicia. Se actualiza mensualmente, por lo que en la actualidad se pueden comprobar los datos de aquellos alojamientos registrados antes del 30/06/2023.

Está compuesto por 18 columnas, pero únicamente serán relevantes para la práctica:

- *codigo_recurso* (clave primaria)
- *municipio*
- *tipo* (hotel, pensión, vivienda de uso turístico…)

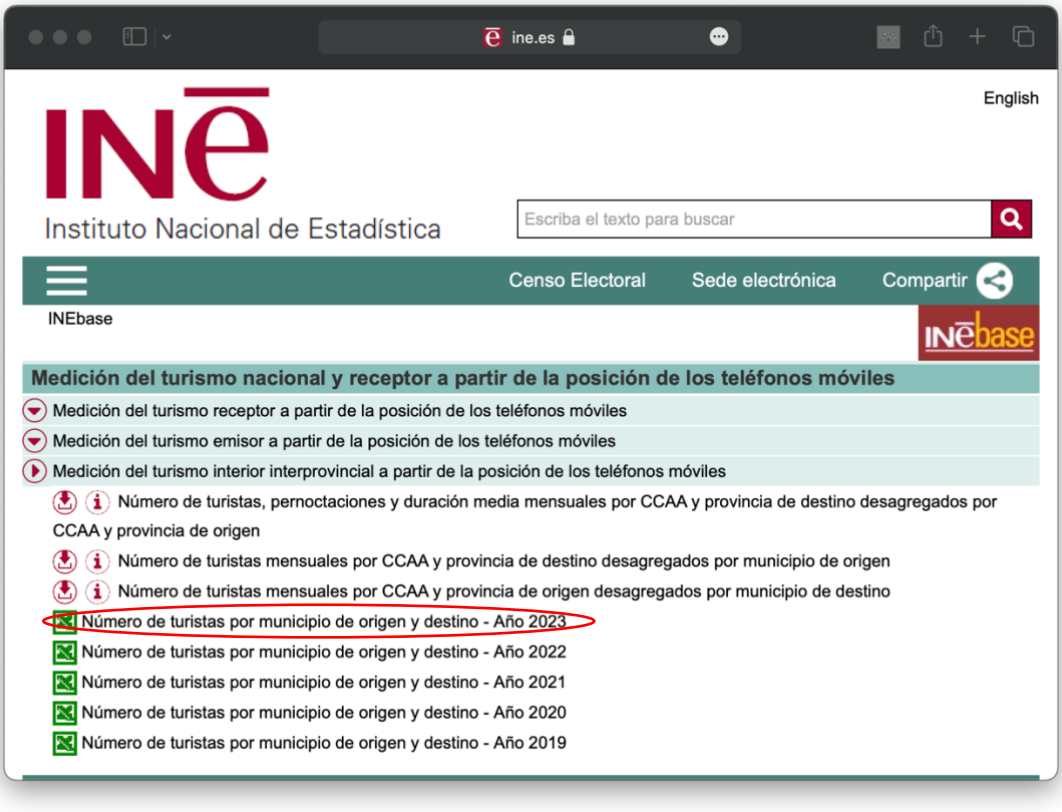
2. **Turismo interior interprovincial**

- nombre del archivo: *exp_tmov_interno_mun_2023.xlsx*

- formato: EXCEL

- fecha: mayo 2023

- url: https://www.ine.es/dynt3/inebase/es/index.htm?padre=8578&capsel=8579



Procedencia del archivo con los datos de turismo interior interprovincial

Este archivo contiene la medición del turismo interior a partir de la posición de los teléfonos móviles. De esta manera, se obtiene una aproximación del número de turistas nacionales que ha recibido cada municipio de España. Sin embargo, se trata de una estadística experimental del INE.

El fichero está formado por varias hojas, pero solo se tendrá en cuenta la correspondiente a mayo de 2023. Cada hoja está compuesta por 10 columnas, pero únicamente serán relevantes para la práctica:

- mun_orig (municipio de origen)
- *dest* (municipio de destino)
- *prov_dest (provincia de destino)*
- *turistas*

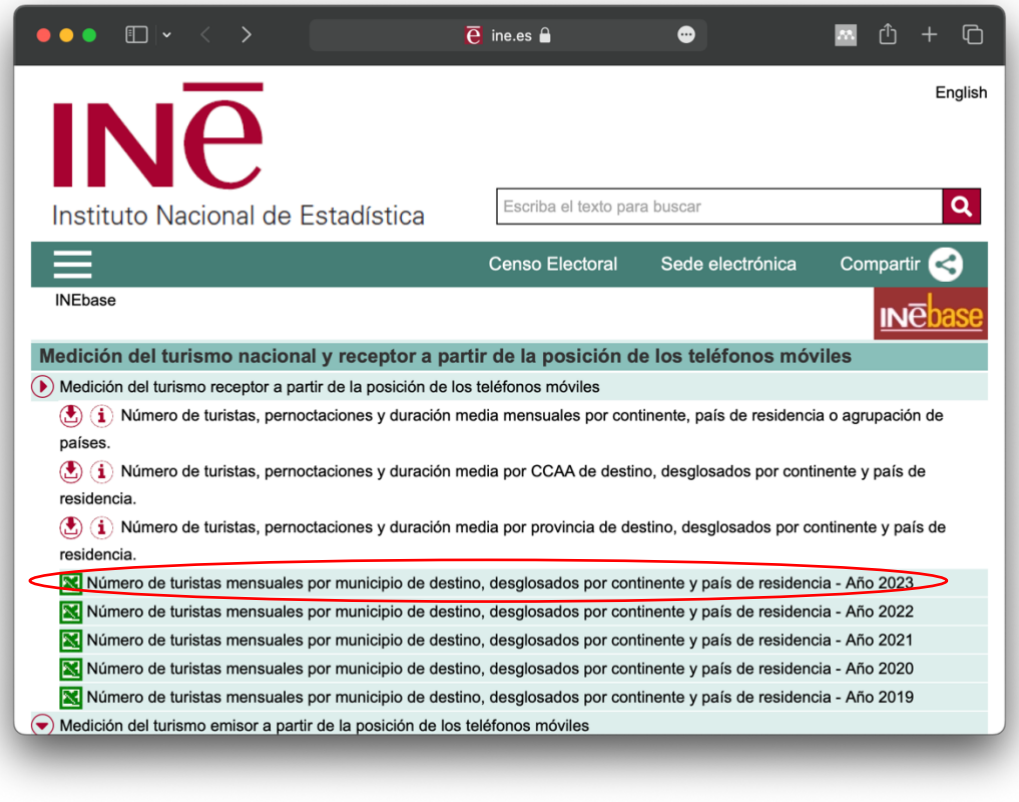
3. **Turismo receptor**

- nombre del archivo: *exp_tmov_receptor_mun_2023.xlsx*

- formato: EXCEL

- fecha: mayo 2023

- url: https://www.ine.es/dynt3/inebase/es/index.htm?padre=8578&capsel=8579



Procedencia del archivo con los datos de turismo receptor

Este archivo contiene la medición del turismo receptor a partir de la posición de los teléfonos móviles. De esta manera, se obtiene una aproximación del número de turistas internacionales que ha recibido cada municipio de España. Sin embargo, se trata de una estadística experimental del INE.

El fichero está formado por varias hojas, pero solo se tendrá en cuenta la correspondiente a mayo de 2023. Cada hoja está compuesta por 8 columnas, pero únicamente serán relevantes para la práctica:

- pais_orig (pais de origen)
- *mun_dest* (municipio de destino)
- *prov_dest (provincia de destino)*
- *turistas*

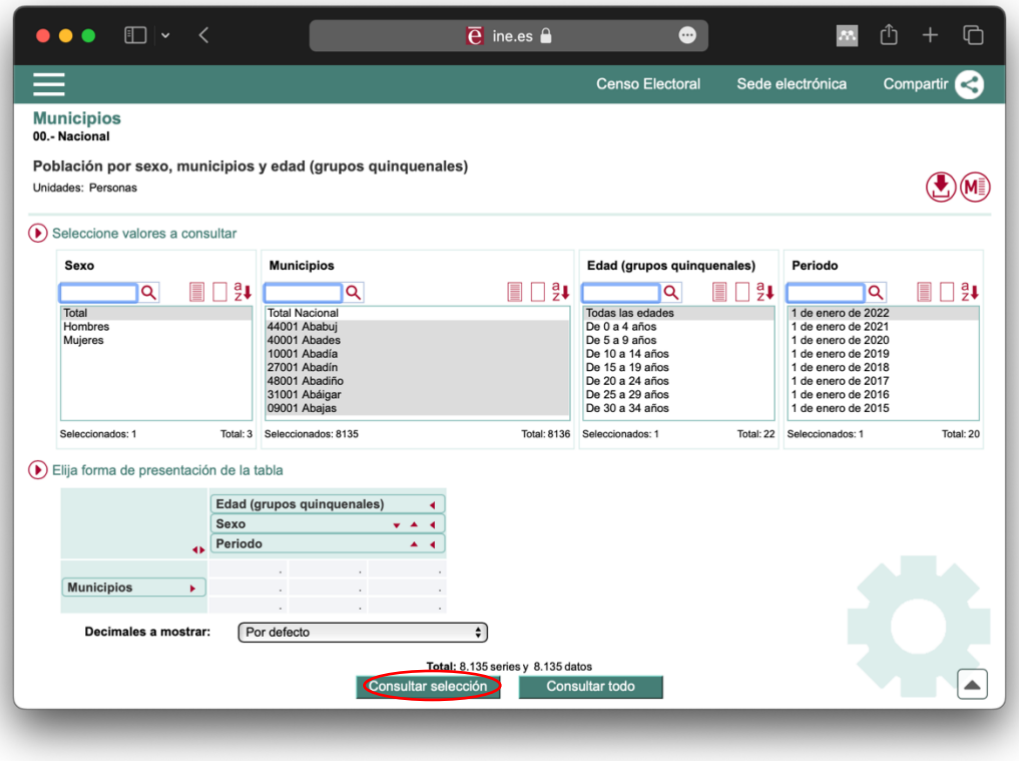
4. **Padrón**

- nombre del archivo: *33570bsc.csv*

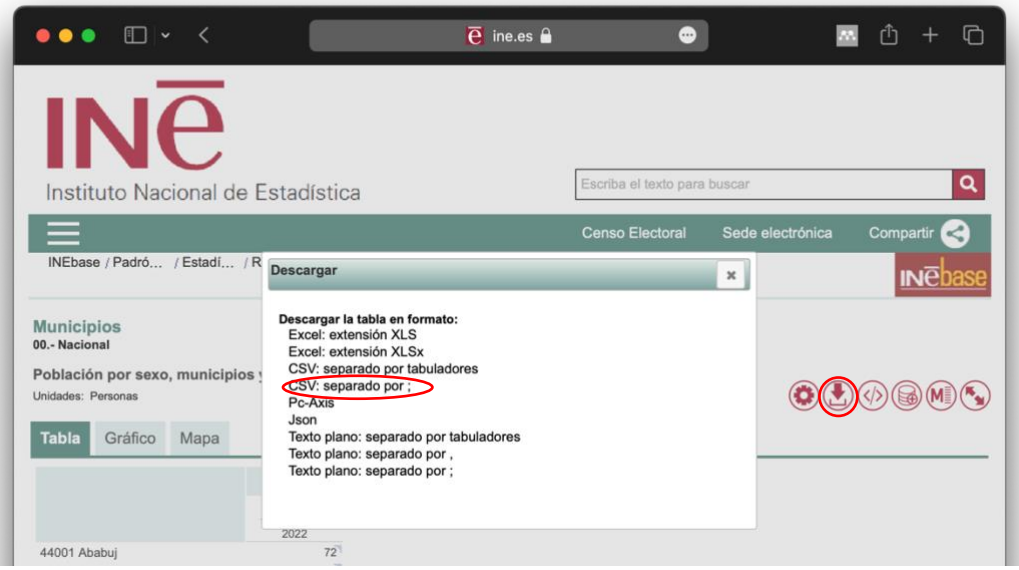
- formato: CSV

- fecha: 1/1/2022

- url: https://www.ine.es/jaxiT3/Tabla.htm?t=33570&L=0



Selección de las variables de la consulta y forma de presentación de la tabla



Procedencia del archivo con los datos de población

Este archivo muestra el número de habitantes de cada municipio de España. Está formado por 5 columnas, pero sólo son relevantes:

- municipios (código del municipio + nombre)
- total (número de habitantes)

Sin embargo, para la resolución de esta práctica no se ha trabajado directamente sobre el archivo CSV, si no que se ha cargado este fichero en una base de datos local de la misma manera que se hizo en la práctica grupal a lo largo del curso.

En primer lugar, se crea la base de datos y el usuario que accederá a ella:

```
mysql -u root -p
        CREATE DATABASE padron;
        CREATE USER ‘user’@’localhost’ IDENTIFIED BY ‘user’;
        GRANT ALL PRIVILEGES ON padron.* to ‘user’@’localhost’;
```

Después, se define la estructura de la tabla padron_2022:

```
mysql -u user –-password=user padron
        DROP TABLE IF EXISTS padron_2022;
        CREATE TABLE padron_2022 (
        municipio VARRCHAR(500),
        edad VARRCHAR(500),
        sexo VARRCHAR(500),
        fecha VARRCHAR(500),
        habitantes VARRCHAR(500) );
```

Finalmente, se carga el fichero de datos. Se utiliza el formato latin1 para evitar problemas con los acentos y se ignora la primera fila puesto que son los nombres de las columnas:

```
mysql -u root -p
        DELETE FROM padron.padron_2022;
        LOAD DATA INFILE 'C:/ProgramData/MySQL/MySOL Server 8.0/Uploads/33570bsc.csv'
        INTO TABLE padron.padron_2022
        CHARACTER SET LATIN1
        FIELDS TERMINATED BY ‘;’ OPTIONALLY ENCLOSED BY ‘”’
        LINES TERMINATED BY ‘\r\n’
        IGNORE 1 ROWS;
```

Ataccama DQ Analyzer

En el análisis de calidad de datos, se estudia en primer lugar cada fuente de datos por separado. Posteriormente, se analizarán las combinaciones que se tendrán que hacer para resolver el problema: turismo nacional y turismo internacional, alojamientos y padron y alojamientos y turismo nacional.

1. **Alojamientos turísticos en Galicia**

Expression	Type	Domain	Non-null	Null	Unique	Distinct	Min	Median	Max
signatura	STRING	pattern	23.784	0	23.705	23.706	""	"VUT-LU-00...	"VUT-SC-00...
_codigo_rec...	STRING	datetime pa...	23.784	0	23.784	23.784	"10"	"200723000...	"VT-PO-000...
_denominaci...	STRING		23.784	0	18.992	20.649	" 1B MONT...	"ISABEL"	"ZÚÑIGA"
direccion	STRING		23.784	0	22.624	22.853	" Nº 1 LETR...	"PR. ENRIQ...	"ZUMALAC...
parroquia	STRING		23.784	0	618	1.399	""	""	"ZORELLE (S...
lugar	STRING		23.784	0	1.150	1.835	""	""	"ZANFOGA"
_codigo_pos...	STRING	pattern	23.784	0	395	1.324	".36900"	"27861"	"PONTE"
municipio	STRING	pattern	23.784	0	11	306	"A ARNOIA"	"O GROVE"	"ZAS"
provincia	STRING	enum pattern	23.784	0	0	4	"A CORUÑA"	"OURENSE"	"PONTEVED...
telefono	STRING	specval	23.784	0	2.441	2.550	" 981 977 2...	""	"988684354"
tipo	STRING	enum pattern	23.784	0	0	8	"ALBERGUE...	"VIVIENDAS...	"VIVIENDAS...
categoria	STRING	enum	23.784	0	0	20	" 1 LLAVE"	""	"GRUPO-D (...
modalidad	STRING	enum pattern	23.784	0	2	10	""	""	"PENSIÓN"
_especialida...	STRING	pattern	23.784	0	13	32	""	""	"XACOBEO"
_habitacione...	INTEGER		23.217	567	45	154	0	3	507
plazas	INTEGER		23.444	340	94	255	1	6	1.913
latitud	FLOAT		20.915	2.869	10.642	13.023	-10	424.000.970...	437.728.143...
longitud	FLOAT		20.915	2.869	10.673	13.040	-927.933.74...	-829.871.31...	-10

Se puede ver como el campo *municipio* presenta una gran cantidad de duplicados, ya que cada municipio puede tener varios alojamientos turísticos. Sin embargo, la estadística más relevante es el número de municipios diferentes (306), lo que indica que, de los 313 municipios gallegos, existen 7 que no han registrado ningún tipo de alojamiento turístico.

Name	Expression	Unique	Non-unique	Null
Code key	_codigo_recurso_	23.784	0	0
Natural key	_denominacion_ + _municipio_ + _direccion_ + _tipo_	23.722	62	0

Además, se comprueba que el campo *codigo_recurso* se puede utilizar como clave primaria, a diferencia de la clave natural que contiene algún duplicado.

Finalmente, se pueden obtener mediciones interesantes examinando el análisis de frecuencias. En el caso del campo *tipo*, se observa que las VUT predominan en Galicia.

Frequency Analysis

Range: none

Value	Count	%
"VIVIENDAS DE USO...	19.663	82,67%
"PENSIONES"	1.371	5,76%
"HOTEL"	944	3,97%
"TURISMO RURAL"	538	2,26%
"ALBERGUES TURÍST...	489	2,06%
"APARTAMENTOS"	407	1,71%
"VIVIENDAS TURÍSTI...	219	0,92%
"CAMPING"	153	0,64%

2. **Turismo interior interprovincial**

Expression	Type	Domain	Non-null	Null	Unique	Distinct	Min	Median	Max
mes	STRING	day enum p...	75.833	0	0	1	2023-05	2023-05	2023-05
mun_orig_cod	STRING	integer patt...	75.833	0	1.317	4.342	01001	28065	52001
mun_orig	STRING		75.833	0	1.313	4.338	Abades	Madrid	Zurgena
dest_cod	STRING	integer patt...	75.833	0	1.440	5.825	01002	26084	52001
dest	STRING		75.833	0	1.437	5.814	Abades	Madrid	Zurgena
turistas	INTEGER		75.833	0	863	1.867	30	52	63.142
prov_orig_cod	STRING	integer patt...	75.833	0	0	52	01	28	52
prov_orig	STRING	pattern	75.833	0	0	52	Albacete	Lugo	Zaragoza
prov_dest_c...	STRING	integer patt...	75.833	0	0	52	01	26	52
prov_dest	STRING	pattern	75.833	0	0	52	Albacete	Huesca	Zaragoza

Se puede ver como los campos *mun_orig* y *dest* presentan una gran cantidad de duplicados, ya que cada municipio puede ser destino de múltiples municipios y viceversa. Se observa también que el número de provincias cuadra con el número de provincias de España (50) más las dos ciudades autónomas.

Name	Expression	Unique	Non-unique	Null
Code key	mun_orig_cod + dest_cod	75.833	0	0
Natural key	mun_orig + dest	75.801	32	0

Además, se comprueba que los campos *mun_orig_cod* y *dest_cod* se pueden utilizar como clave primaria. Sin embargo, si simplemente se utilizan los nombres de los municipios de origen y destino, hay duplicados. Este caso aparecerá en el análisis del padrón, donde se descubre que hay varios municipios en España con el mismo nombre.

3. **Turismo receptor**

Expression	Type	Domain	Non-null	Null	Unique	Distinct	Min	Median	Max
mes	STRING	day enum p...	32.965	0	0	1	2023-05	2023-05	2023-05
pais_orig_cod	STRING	integer patt...	32.965	0	9	130	000	102	504
pais_orig	STRING	pattern	32.965	0	9	130	Albania	Suiza	Vietnan
mun_dest_c...	STRING	integer patt...	32.965	0	245	4.780	01001	23067	52001
mun_dest	STRING		32.965	0	245	4.773	Abades	Manilva	Zurgena
turistas	INTEGER		32.965	0	1.790	3.177	30	93	711.889
prov_dest_c...	STRING	integer patt...	32.965	0	0	52	01	23	52
prov_dest	STRING	pattern	32.965	0	0	52	Albacete	Guadalajara	Zaragoza

Name	Expression	Unique	Non-unique	Null
Code Key	pais_orig_cod + mun_dest_cod	32.965	0	0
Natural Key	pais_orig + mun_dest	32.893	72	0

Se observan las mismas características que en el apartado anterior. Sin embargo, hay que tener en cuenta para pasos posteriores que el campo *país_orig* cuenta con el valor *TOTAL*, que será el que se utilice para los cálculos.

4. **Padrón**

Expression	Type	Domain	Non-null	Null	Unique	Distinct	Min	Median	Max
municipio	STRING		8.135	0	8.135	8.135	01001 Alegr...	26017 Arne...	52001 Melilla
edad	STRING	enum pattern	8.135	0	0	1	Todas las ed...	Todas las ed...	Todas las ed...
sexo	STRING	enum pattern	8.135	0	0	1	Total	Total	Total
fecha	STRING	day enum p...	8.135	0	0	1	1 de enero ...	1 de enero ...	1 de enero ...
habitantes	STRING	float pattern	8.131	4	2.406	3.572	1.000	3.386	999
substr(municipio, indexOf(municipio, ' ')+1)	STRING		8.135	0	8.101	8.118	Ababuj	Mascaraque	Zurgena
substr(municipio, 0, indexOf(municipio, ' '))	STRING	integer patt...	8.135	0	8.135	8.135	01001	26017	52001

En este caso, hay que tener en cuenta que el campo *municipio* está realmente formado por: código + “ ” + nombre. Por lo que en esta combinación no se encuentra ningún duplicado y se puede usar como clave primaria. Sin embargo, si se separa este campo en dos, se observa como el nombre presenta duplicados.

Esto queda todavía más claro si se observan las frecuencias. Se puede ver como 34 municipios de España tienen un nombre que no es único.

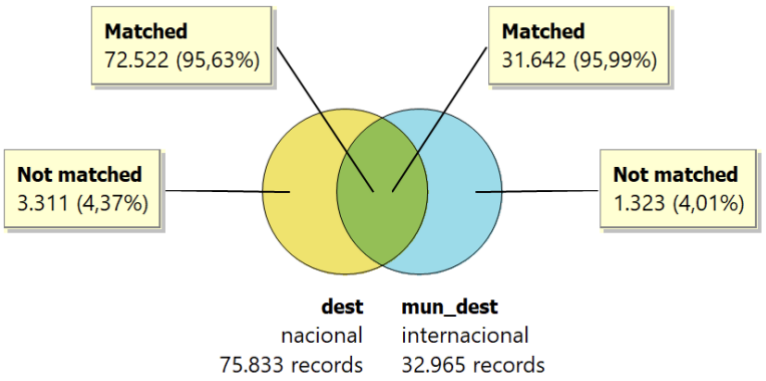
100 most common values:

Value	Count	%
Arroyomolinos	2	0,02%
Cabanes	2	0,02%
Campillo, El	2	0,02%
Castejón	2	0,02%
Cieza	2	0,02%
Fonfría	2	0,02%
Mieres	2	0,02%
Molar, El	2	0,02%
Moya	2	0,02%
Rebollar	2	0,02%
Sada	2	0,02%
Sancti-Spíritus	2	0,02%
Sobrado	2	0,02%
Torrent	2	0,02%
Villaescusa	2	0,02%
Villanueva de los In...	2	0,02%
Zarza, La	2	0,02%
Ababuj	1	0,01%

Será especialmente relevante el caso de Sada y Sobrado, puesto que cuando se junte el padrón con los alojamientos de Galicia por el nombre del municipio, habrá conflictos. Es por ello por lo que se usará el código del municipio para filtrar aquellos que empiecen por 15, 36, 27 o 32 (códigos de Galicia).

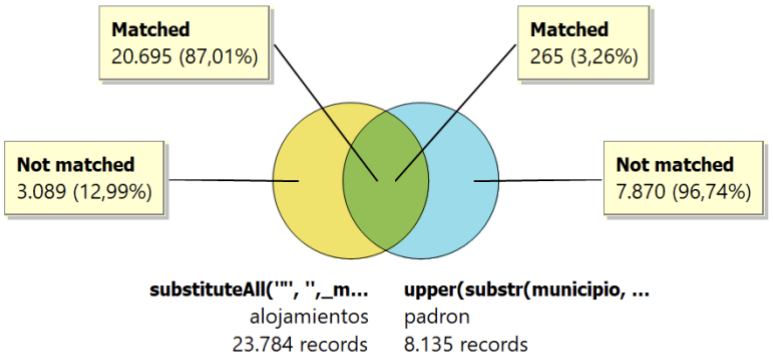
Otra peculiaridad de esta base de datos es que hay 4 municipios cuyo número de habitantes es nulo. Se eliminarán estos casos, puesto que corresponden a municipios que ya no existen.

5. **Turismo nacional y turismo internacional**



En este *join*, se puede ver que existen municipios que han tenido únicamente o turismo nacional o turismo internacional. Sin embargo, se trata de un bajo porcentaje. Posteriormente, en la herramienta Denodo, se comprobará que en Galicia todos los municipios que han recibido turismo internacional han recibido también nacional; por lo que se podrá aplicar un *left-join*.

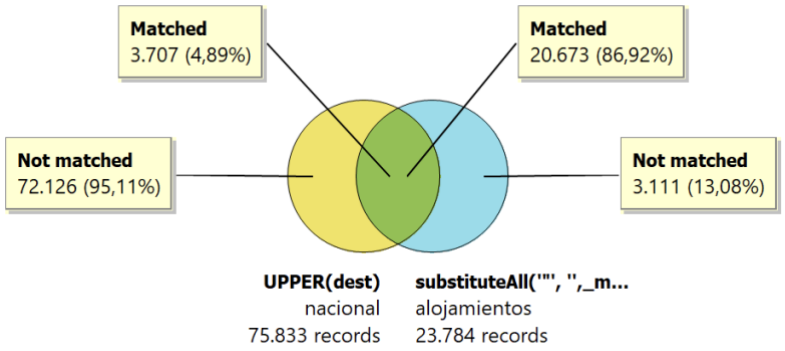
6. **Alojamientos y padrón**



Lo normal en este caso sería que el 100% de los municipios que aparecen en el fichero de alojamientos estuvieran en el padrón. En cambio, se obtiene un porcentaje del 87.01% debido a que en el fichero de alojamientos aparecen municipios como “A Coruña”, “A Baña” y “O Grove”, que en el padrón figuran como “Coruña, A”, “Baña, A” y “Grove, O”. Este problema se resolverá posteriormente en Denodo.

Por otro lado, el 3.26% de match por parte del padrón se debe a que el fichero de alojamientos solo tiene los municipios gallegos, y el padrón es a nivel nacional (a parte del problema anterior).

7. **Turismo nacional y alojamientos**



De forma similar al caso anterior, el 86.92% de match por parte del fichero de alojamientos se debe principalmente al diferente formato de los nombres como “A Coruña”. Sin embargo, también se puede deber en menor medida a municipios que tienen alojamientos turísticos, pero no han recibido ningún turista.

Igual que antes, el 4.89% de match por parte del turismo nacional se debe a que el fichero de alojamientos solo tiene los municipios gallegos, y el turismo nacional contempla el turismo en todos los municipios de España.

Se analiza este *join*, puesto que alojamientos con padrón y nacional con internacional se resolverán ambos con un *left-join*.

Denodo Express

1. **Data sources**

Para importar los datos de los **alojamientos**, se selecciona “delimited file”, se especifica la ruta local del archivo, “;” como delimitador de columnas, “\n” como delimitador de filas y se indica que se omitan las 5 primeras filas porque no son relevantes. Además, se marca que tiene cabecera.

En cuanto a los datos de **turismo nacional**, se selecciona “excel”, se especifica la ruta local del archivo y se indica que únicamente utilice la hoja “2023-05”, correspondiente a los datos de mayo. Además, se marca que tiene cabecera.

Para los datos de **turismo internacional**, se selecciona “excel”, se especifica la ruta local del archivo y se indica que únicamente utilice la hoja “m05_2023”, correspondiente a los datos de mayo. Además, se marca que tiene cabecera.

Finalmente, para la base de datos de **padrón**, se selecciona “JDBC”, utilizando el adaptador “MySQL 5” y la URL “jdbc:mysql://localhost:3306/padron”. El usuario es *user* y la contraseña *user*.

2. **Base views**

Se crean todas las vistas base de forma sencilla. En este paso se definen las claves primarias de cada fuente, así como el tipo de datos de las columnas relevantes.

3. **Derived views**

3.1. *vt_por_municipio*

A partir de *alojamientos*, se seleccionan aquellas filas donde el tipo de alojamiento sea “VIVIENDAS DE USO TURÍSTICO" y se agrupa por el nombre del municipio. El resultado muestra el nombre del municipio y el número de viviendas de uso turístico que tiene.

23.784 filas x 18 columnas → 287 filas x 2 columnas

3.2 *padron_mun_cod_separado*

Mediante el uso de las *substring*() y *position*(), se separa el campo municipio, en municipio (nombre) y código. Además, se elimina el punto que separa los miles de las cifras de habitantes, para poder pasarlo a entero.

8.135 filas x 5 columnas → 8.135 filas x 3 columnas

3.3 *padron_arreglado*

Se seleccionan los municipios cuyo código empieza por 15, 36, 27 o 32 (códigos de Galicia). Además, se eliminan aquellos municipios donde el número de habitantes es nulo. Finalmente, con las funciones *substring*() y *position*() se consigue pasar de nombres de municipios como “Coruña, A” a “A Coruña”. Esto facilitará el posterior *join*.

```
case WHEN (municipio like '%,%')

THEN concat(substring(municipio, (position(',' IN municipio)+1), len(municipio)), ' ', substring(municipio, 0, (position(',' IN municipio)-1)))

ELSE municipio

END
```

8.135 filas x 3 columnas → 313 filas x 2 columnas

3.4 *vt_cada_100_hab*

Se añade a la vista vt_por_municipio el número de habitantes de sus municipios y se calcula cuantas viviendas de uso turístico tienen por cada 100 habitantes.

287 filas x 2 columnas ∞ 313 filas x 2 columnas → 287 filas x 4 columnas

3.5 *top_municipios*

Se filtra la vista anterior para quedarse solo con los municipios que tengan al menos 1 vivienda de uso turístico por cada 100 habitantes.

287 filas x 4 columnas → 61 filas x 4 columnas

3.6 *nacional_solo_galicia*

Se mantienen solo aquellos registros donde la provincia de destino sea gallega. Se agrupa por destino y se suman los turistas para obtener los turistas nacionales totales por cada municipio.

75.833 filas x 10 columnas → 308 filas x 2 columnas

3.7 inter*nacional_solo_galicia*

Se mantienen solo aquellos registros donde la provincia de destino sea gallega y el país de origen sea "Total".

32.965 filas x 8 columnas → 288 filas x 2 columnas

3.8 *turismo_total_por_municipio*

Para aquellos municipios que han tenido turismo nacional, se observa si han tenido turismo internacional. En caso de ser nulo, se cuenta como 0.

Además, igual que la vista 3.3, se pasa de nombres de municipios como "Coruña, A" a "A Coruña".

308 filas x 2 columnas ∞ 288 filas x 2 columnas → 308 filas x 3 columnas

3.9 *vista_final*

Para los municipios de *top_municipios*, se le añaden la cantidad de turistas nacionales, internacionles y totales.

61 filas x 4 columnas ∞ 308 filas x 3 columnas → 61 filas x 7 columnas

4. **Data services**

Por último, se crea un servicio REST para acceder a los datos de la vista_final.

municipio	numero_de_vt	habitantes	vt_cada_100_hab	turistas_nacional_mayo_2023	turistas_internacional_mayo_2023	turistas_total_mayo_2023
Sanxenxo	2624	17760	14.8	17740	2986	20726
Vilagarcía de Arousa	359	37677	1	11868	797	12665
O Grove	606	10809	5.6	9714	1065	10779
Boiro	231	18976	1.2	7724	460	8184
Ribeira	292	26897	1.1	6613	686	7299
Ribadeo	241	9811	2.5	6733	563	7296
Nigrán	290	18054	1.6	3821	1899	5720
Cangas	577	26832	2.2	4310	1136	5446
Viveiro	342	15231	2.2	4502	382	4884
Baiona	342	12349	2.8	3296	1532	4828