Mario Subías Pérez y María Nieves Pérez Gil

PRÁTICA 1 – WEB SCRAPING

1. Contexto. Explicar en qué contexto se ha recolectado la información. Explique por qué el sitio web elegido proporciona dicha información.

El Instituto Nacional de Estadística (INE) de España es el encargado de coordinar servicios estadísticos y la vigilancia, control y supervisión de los procedimientos técnicos. Es un organismo autónomo donde se destacan las estadísticas económicas, demográficas y sociales españolas que son públicas a través de tu página web oficial.

De entre todas las estadísticas que se muestran, hemos decidido recolectar información sobre las cifras de población en España por Comunidades Autónomas y Provincias. Estas cifras nos servirán para hacer pirámides poblacionales o comparaciones de censos por fechas, grupos de edad, sexo o lugar.

La pirámide de población representa gráficamente las características de una población perteneciente a una localidad o país en un momento concreto del tiempo empleando rango quinquenales generalmente. Esto nos permitirá ver como es la distribución y comparar distintos momentos o distintos lugares.

2. Definir un título para el dataset. Elegir un título que sea descriptivo.

El dataset va a llamarse "Población de España por grupos quinquenales especificando país de nacimiento y sexo."

3. Descripción del dataset. Desarrollar una descripción breve del conjunto de datos que se ha extraído (es necesario que esta descripción tenga sentido con el título elegido).

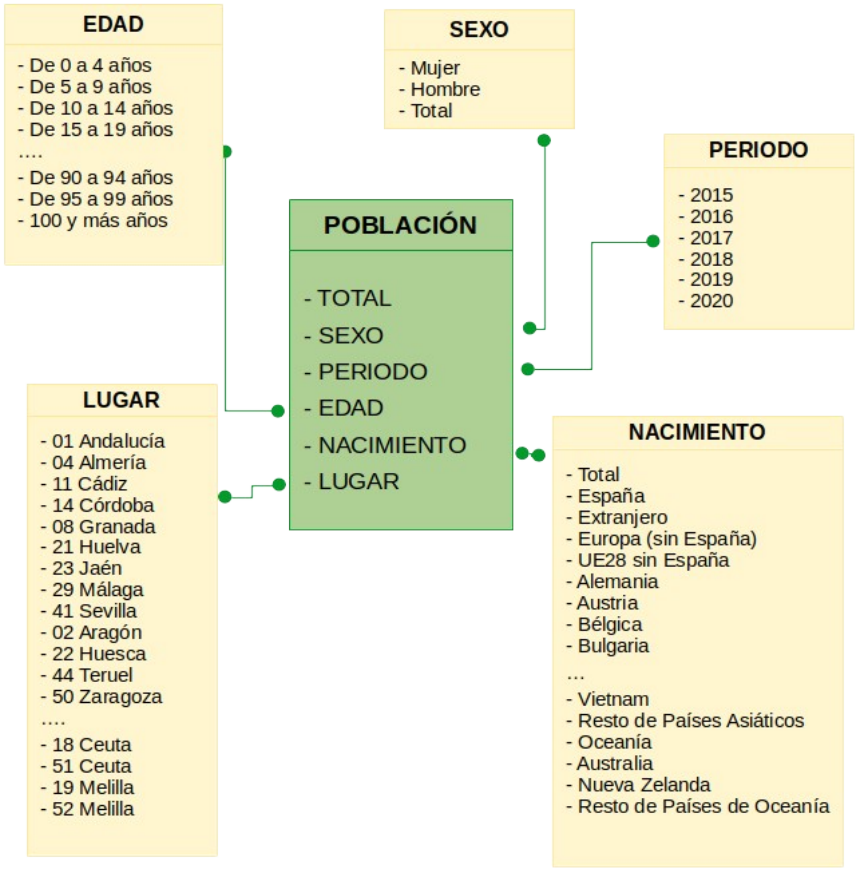
El dataset tendrá las cifras de población de las comunidades autónomas y provincias españolas, junto con la total. Los campos que lo forman son el lugar (comunidad autónoma/provincia) de la que se va a medir la población, el grupo quinquenal de edades, el periodo donde se han medido los datos, el sexo, el país de nacimiento y la población que corresponda.

Por cada comunidad autónoma y provincia se muestran los datos de población desglosados por los distintos países de nacimiento, es decir, se muestran de una forma específica por cada país y una forma un poco más genérica como por ejemplo por el término "extranjero" o "Union europea sin España".

4. Representación gráfica. Presentar esquema o diagrama que identifique el dataset visualmente y el proyecto elegido

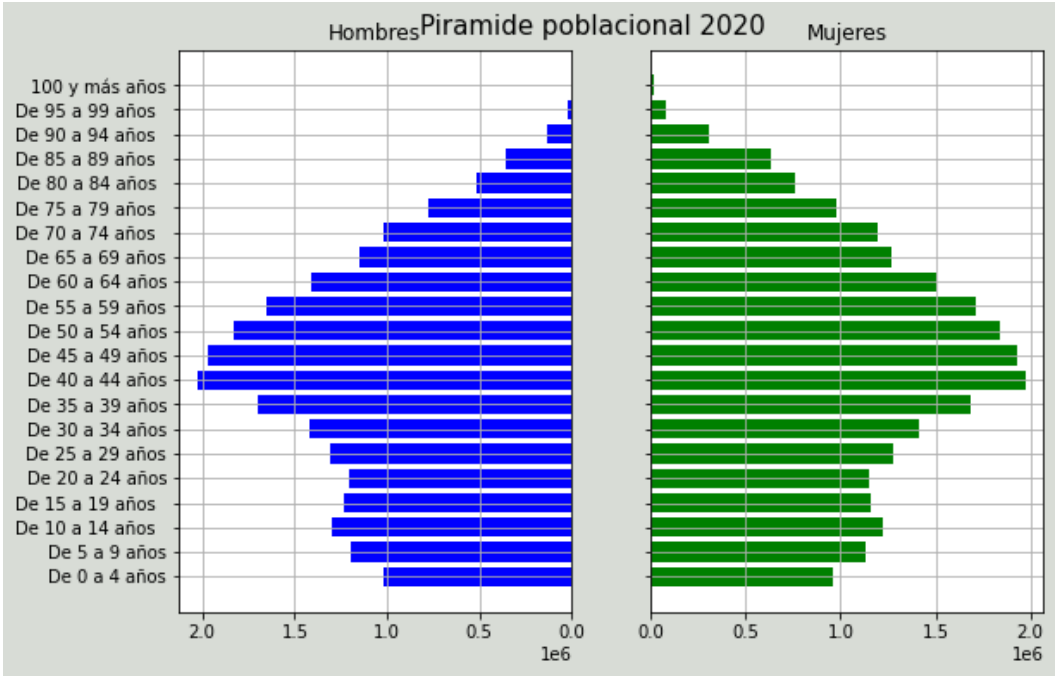
Para que se vea de una forma más clara como está compuesto el dataset y ver sus campos de una forma más visual se presenta el siguiente esquema:

Mario Subías Pérez y María Nieves Pérez Gil



Cada fila del csv obtenido corresponde con el siguiente esquema, unos datos de población (correspondientes al valor *Total*) para un grupo de edad, un periodo, para un país correspondiente al país de nacimiento, sexo y el lugar donde se quiere consultar la población.

De esta forma, se pueden pueden consultar datos y generar gráficos de comparatívas entre lugares o países, o por ejemplo pirámides de población como la siguiente:



Mario Subías Pérez y María Nieves Pérez Gil

Dibujada con el dataset que hemos capturado, para el entorno nacional, todos los grupos de edad y cualquier país de nacimiento para el periodo 2020, código que aparece en population.

5. Contenido. Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se ha recogido.

Explicando más a fondo, los campos que incluye el dataset son:

- Lugar: Un campo de tipo String que contiene tantos valores como comunidades autónomas y provincias hay en España, junto con una opción de total, que correspondería a la suma de los valores de población de las comunidades autónomas en condiciones de total.

- Nacimiento: Corresponde al país de nacimiento de la población que se indique, es decir, para cada comunidad y provincia, el país de nacimio de la población que indique en el grupo de edad correspondiente.

- Edad: Grupos quinquenales que toma valores cada 4 años, empezando por “De 0 a 4 años” y acabando en “De 95 a 99 años” y “100 y más años”.

- Sexo: Valor categórico que vale “hombre” o “mujer” .

- Periodo: Valor de tipo entero que corresponde al periodo donde se han capturado los datos. En este caso tenemos valores de 2015 a 2020 incluídos.

- Total: Valor de tipo entero que corresponde a la población.

En la siguiente imagen se puede ver unas líneas del fichero csv formado con los datos anteriores:

Para recoger estos valores, se ha realizado una petición a la página oficial del INE que ofrece datos detallados sobre por tablas agrupadas por comunidades y provincias, en este caso de censos (https://www.ine.es/dynt3/inebase/index.htm?padre=6235&capsel=6692).

Toda la información que se muestra es semejante pero agrupada y mostrada de diferente manera. Hemos decidido que la opción de "Población por país de nacimiento, edad (grupos quinquenales) y sexo" es completa y nos da más información que en otros casos, como el país de nacimiento.

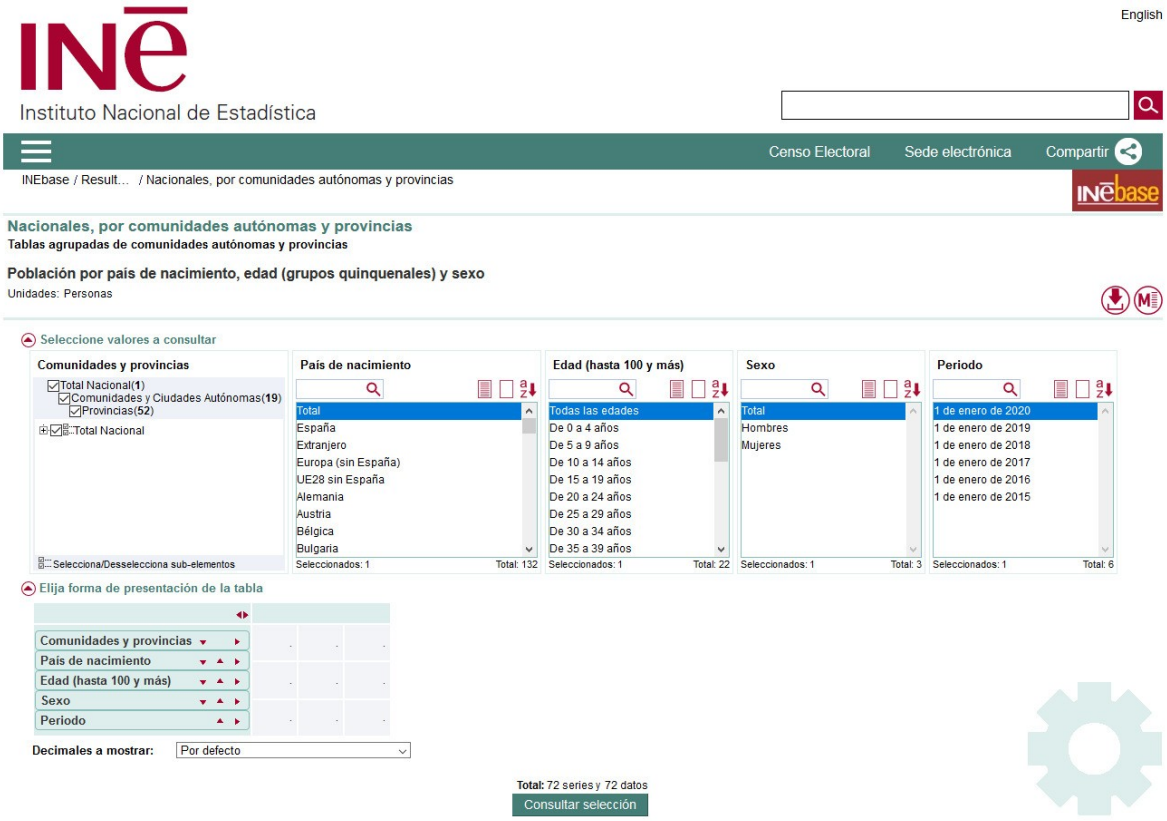
Accediendo a dicha opción mediante el método get de requests, se realiza otra petición para obtener el set de datos correspondiente de la forma que nos conviene, en este caso separando los valores por ';'. Existe otra opción que se muestran los datos en tablas pero no es posible su recolección mediante esta técnica ya que se carga con javascript y requests no es capaz de leerlo.

Una vez tenemos los datos separados por ';', resultado de hacer una petición a la url que corresponde, se cargan en un dataframe para que sea más fácil su parseo y filtrado. Se eliminan filas que se pueden obtener desde otra forma o filtrado para así no tener datos duplicados y se modifican los tipos de datos de alguna de las columnas para que sea más fácil cuando posteriormente se quiera obtener información.

6. Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de análisis anteriores o, en caso de no haberlas, justificar esta búsqueda con análisis similares.

Gracias a la base de datos pública del INE, Instituto Nacional de Estadística, se pueden obtener tantos datos como se quiera para poder obtener conclusiones sobre temas demográficos, sociales o económicos de España.

En nuestro caso se han obtenido datos de la siguiente web correspondiente a datos censales agrupados por comunidad autónoma y provincias esoañolas:



Mario Subías Pérez y María Nieves Pérez Gil

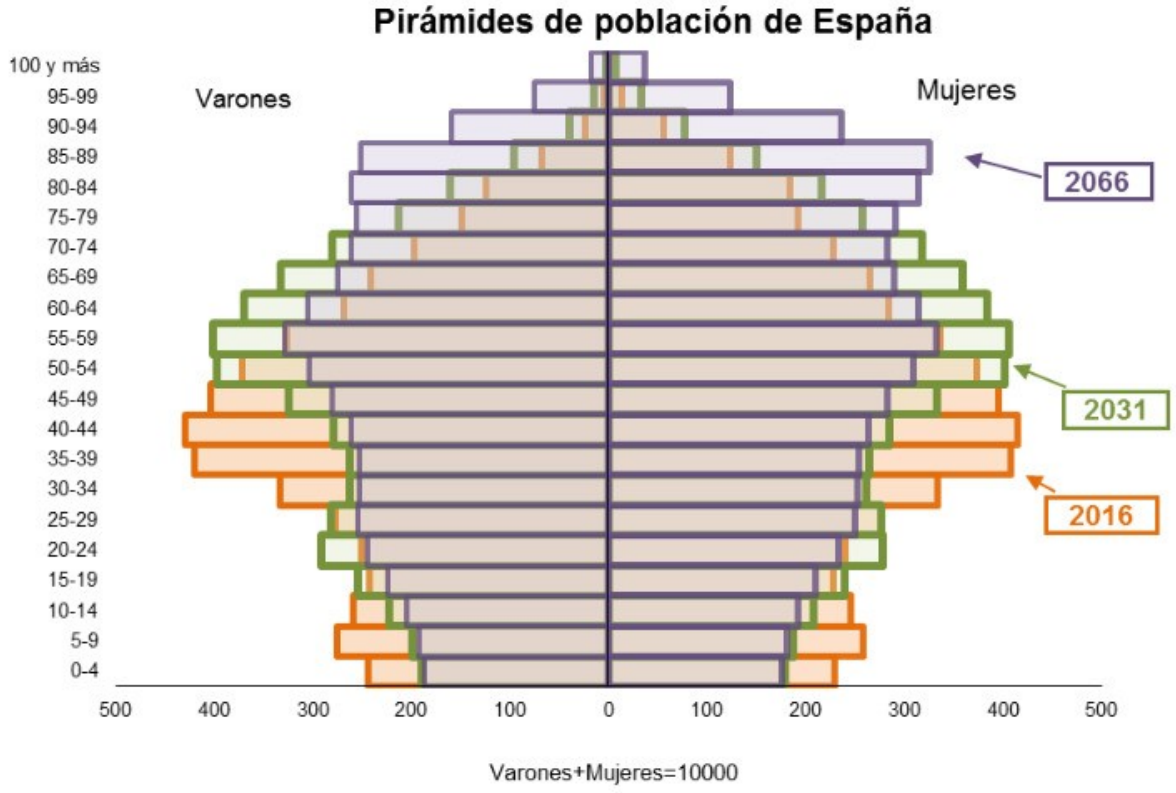
Existen análisis similares al nuestro que hablan sobre las pirámides de población como son:
https://www.defensordelpueblo.es/wp-content/uploads/2019/06/Separata_situacion_demografica.pdf
https://www.ine.es/infografias/infografia_dia_poblacion.pdf
https://www.ine.es/prensa/np994.pdf

7. Inspiración. Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder. Es necesario comparar con los análisis anteriores presentados en el apartado 6.

Desde la página oficial del INE se ofrencen variedad de datos públicos, como los elegidos por nosotros en este caso, población por país de nacimiento y sexo, se pueden crear variedad de esquemas y diagramas para ver como ha ido variando un territorio cuando hablamos de censos.

Por ejemplo, este dataset permitirá hacer pirámides de poblacion de distintos lugares y comparativas en cuanto a años. Una pirámide de población es una representación de los datos de la población de un país basados en sexo y edad que permite comparaciones de una forma rápida entre lugares o periodos de tiempo, que permite ver fenómenos como el envejecimiento o equilibrio de sexos.

Se puede ver como en los artículos citados en el apartado anterior, también se hacen análisis demográficos a través de pirámides de población como se en nuestro caso y donde se puede ver a continuación en la imagen obtenida en uno de los análisis previos:



Además de pirámides de población, se puede comparar como ha ido variando la población en varios periodos de sitios concretos, medir densidades de población en distintos momentos y lugares o medir el porcentaje de población nacional y extranjera, además de ver su país de nacimiento para ver correlaciones entre países y lugares o grupos poblacionales de edad.

8. Licencia. Seleccione una de estas licencias para su dataset y explique el motivo de su selección:
○ **Released Under CC0: Public Domain License**
○ **Released Under CC BY-NC-SA 4.0 License**
○ **Released Under CC BY-SA 4.0 License**
○ **Database released under Open Database License, individual contents under Database Contents License**
○ **Other (specified above)**
○ **Unknown License**

Según el apartado de aviso legar de la web del INE (www.ine.es):

Reutilización de la información contenida en este sitio web

La información contenida en este sitio web procede de múltiples fuentes, por lo que el INE solo autoriza la reutilización de aquélla cuya fuente original sea el propio INE. Esta reutilización podrá tener objeto comercial o no comercial y se realizará siempre bajo las siguientes condiciones generales:

- Se prohíbe expresamente desnaturalizar el sentido de la información.
- Debe citarse la fuente de la información objeto de reutilización. Esta cita podrá realizarse de la siguiente manera: **Fuente: Sitio web del INE: www.ine.es** si no se realiza ningún tratamiento de los datos o bien: **Elaboración propia con datos extraídos del sitio web del INE: www.ine.es** en caso de que se realice tratamiento de los datos.
- Debe mencionarse la fecha de la última actualización de la información objeto de reutilización, siempre y cuando estuviera incluida en el original.
- No se podrá indicar, insinuar o sugerir que el INE participa, patrocina o apoya la reutilización que se lleve a cabo con la información.
- El INE no será responsable del uso que de su información hagan los agentes reutilizadores. Tampoco será responsable de los daños materiales o sobre datos, ni de posibles perjuicios económicos provocados por el uso de la información reutilizada.

Esta licencia de uso se rige por las leyes españolas independientemente del entorno legal del usuario. Cualquier disputa que pueda surgir en la interpretación de este acuerdo se resolverá en los tribunales españoles.

Cualquier duda o comentario sobre el contenido de este servidor debe dirigirse a la Subdirección General de Difusión Estadística del INE:

*Formulario de consultas: **www.ine.es/infoine**, tlf: (+34) 91 583 91 00; fax: (+34) 91 583 91 58*

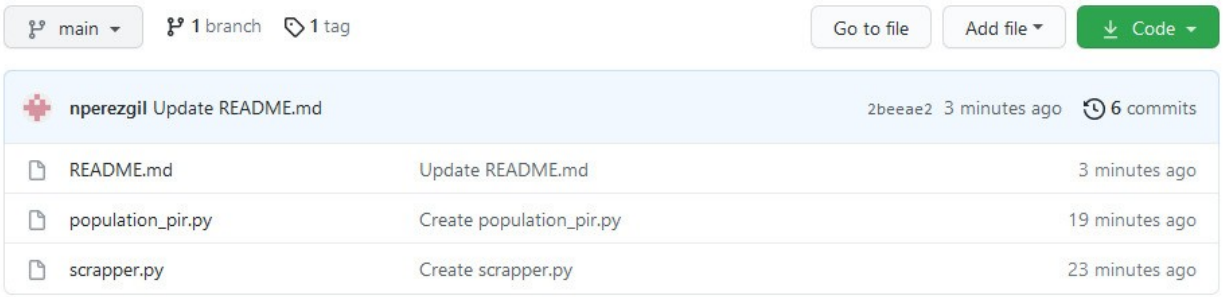
Se puede ver que permite la reutilización de la información contenida en este sitio web siempre que se cumplan las condiciones establecidas en la imagen anterior, que son:

- No desnaturalizar el sentido de la información
- Citar la fuente de la información, en este caso **Fuente: Sitio web del INE: www.ine.es**
- Incluír fecha de la última actualización
- Desvincular al INE en la obtención de la información

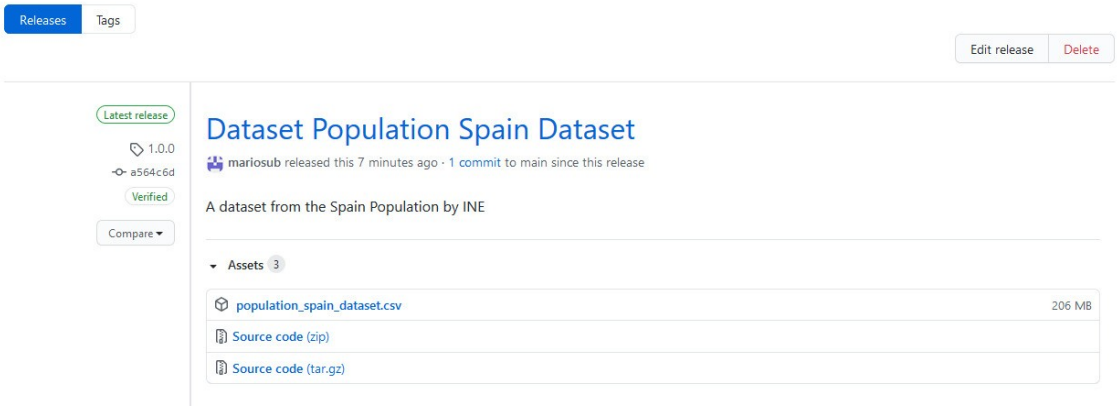
Por tanto, la licencia que utilizaremos será **Released Under CC0: Public Domain License** ya que son datos de dominio público y el INE indica que se podrá tener objeto comercial o no comercial de los mismos.

9. Código. Adjuntar el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R.

Se ha subido a github el código correspondiente al scrapper de los datos del INE en el fichero scrapper.py así como el código de generación de la pirámide poblacional en el fichero population_pir.py.



Se puede ver también como además de los códigos, se ha subido el fichero csv con los datos.

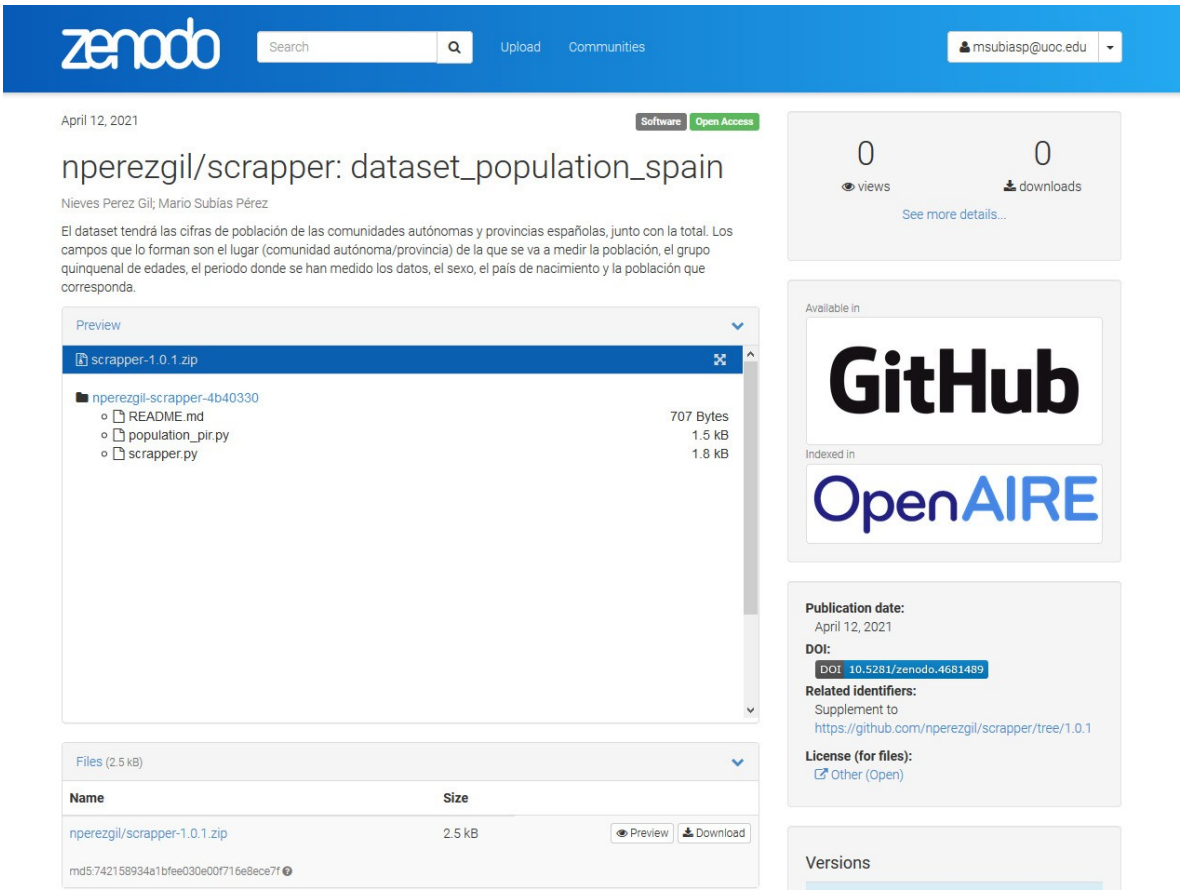


https://github.com/nperezgil/scrapper

10. Dataset. Publicación del dataset en formato CSV en Zenodo (obtención del DOI) con una breve descripción.

Dataset publicado y linkado entre zenodo y github con el siguiente DOI:

https://doi.org/10.5281/zenodo.4681489



Podemos ver que además de haber cargado el dataset, hemos unido la información con los códigos python en github.

Anexo contribuciones.

Investigación previa	M.S.P, M.N.P.G
Redacción de las respuestas	M.S.P, M.N.P.G
Desarrollo código	M.S.P, M.N.P.G