

## PRACTICA 1.

### Dataset de afiliaciones a la Seguridad Social por tipos de régimen y sectores de actividad

*Jordi Sánchez Ferrer y Tania Piñeiro Vidal*

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

La declaración de la pandemia por la Organización Mundial de la Salud en marzo de 2020 ha provocado un shock económico sin precedentes en la economía española. En términos de empleo, las medidas de protección y flexibilización han tendido a contrarrestar la caída, sin embargo, su impacto se ha reflejado en el número de afiliados a la Seguridad Social. Paralelamente, el impacto de la pandemia sobre el mercado de trabajo está siendo desigual, siendo los más afectados las mujeres y los jóvenes. También su efecto ha sido distinto en los diferentes sectores, siendo el sector servicios el que se ha visto más damnificado. Es ahí donde reside la importancia de este proyecto, ya que se pretende estudiar las diferencias en las afiliaciones a la Seguridad Social en cada mes para los diferentes sectores y regímenes.

La afiliación al sistema de la Seguridad Social es obligatoria para todas las personas incluidas en el campo de aplicación de la Seguridad Social y única para toda la vida del trabajador, se organiza en la actualidad en los regímenes siguientes: general, especial de la minería del carbón, especial agrario, especial de empleados del hogar, especial de trabajadores autónomos y especial de trabajadores del mar. Además se reconocen 5 sectores principales: agricultura, industria, construcción, servicios y no clasificables.

El dataset seleccionado recoge información mensual del número de afiliaciones a la Seguridad Social por régimen (régimen general, autónomos y total) y por sectores de actividad (total, agricultura, industria, construcción, servicios y no clasificables) para toda España y para la comunidad de Catalunya. Las series también muestran las variaciones interanuales.

En todos los casos la información procede de la explotación estadística del fichero de afiliación a los diferentes regímenes de la Seguridad Social, cuya gestión corresponde a la Tesorería General de la Seguridad Social.

Debe tenerse en cuenta que se incluyen los afiliados en alta laboral o en situaciones asimiladas (incapacidad temporal, suspensión por regulación de empleo, paro parcial, etc.) y que el número de afiliados no corresponde necesariamente al de trabajadores, sino al de situaciones que generan obligaciones de cotizar: la misma persona se contabiliza tantas veces como situaciones de cotización tenga, ya sea porque tiene varias actividades laborales en un mismo régimen o en otras.

El sitio web del que se obtiene la información es el Instituto de Estadística de Catalunya y la fuente original de los datos es el Ministerio de Trabajo, Migraciones y Seguridad social. El Instituto de Estadística de Cataluña (Idescat) es el organismo oficial de estadística de Cataluña cuya misión es proveer información estadística relevante y de alta calidad, con independencia profesional, y coordinar el Sistema estadístico de Cataluña, con el objetivo de contribuir a la toma de decisiones, la investigación y la mejora de las políticas públicas. La información de los datos de afiliación a la Seguridad Social forma parte de los indicadores de coyuntura económica, en concreto de trabajo, recogidos por Idescat.

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

El título del dataset es “Afilaciones a la Seguridad Social”.

## 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 conjunto de datos generado en esta práctica recoge información sobre el número de afiliaciones a la Seguridad Social y el porcentaje de variación interanual, por régimen y sector, en España y en Catalunya. Estos datos se recogen mensualmente y en el momento actual se dispone de datos desde 01/1994 hasta 02/2021, es decir, más de 27 años.

Nuestro objetivo es que esta información fuese accesible, recogida en un archivo csv y estructurada de tal forma que se pueda filtrar fácilmente la información en función del periodo, región, régimen o sector de interés. Para ello, el dataset se ha estructurado en 7 columnas que se describen a continuación:

1. **Año:** Año al que corresponde la medición de los datos en formato numérico.
2. **Mes:** Mes al que corresponde la medición de los datos en formato numérico.
3. **Región:** Región a la que corresponde la medición de los datos. En el momento actual se dispone de información correspondiente a la comunidad de Cataluña (de forma que aparecerá ‘Cataluña’ en el campo ‘Región’) y también la referente al valor global para toda España (de forma que aparecerá ‘España’ en el campo ‘Región’)
4. **Régimen:** Los regímenes de la Seguridad Social son sistemas establecidos por la Ley para garantizar una base de bienestar y prestaciones sociales a los contribuyentes españoles y que varían en función de la situación del ciudadano a nivel laboral. En este dataset se han agrupado en tres grupos:
  - a. Régimen general y minería del carbón: Todos los trabajadores españoles por cuenta ajena de la industria y los servicios y asimilados a los mismo. El régimen general no incluye el sistema especial agrario ni el sistema especial del hogar.
  - b. Autónomos: Trabajadores por cuenta propia o autónomo son aquellos que realizan de forma habitual, personal y directa una actividad económica a título lucrativo, sin sujeción por ella a contrato de trabajo.
  - c. Total: La afiliación total incluye el régimen general y los especiales de la minería del carbón, agrario, de empleados del hogar, de trabajadores autónomos y de trabajadores del mar.
5. **Sector:** Se contemplan 6 tipos de sectores diferentes en los cuales se agrupan las diferentes actividades laborales: Agricultura, Industria, Construcción, Servicios, No clasificables (que no se pueden atribuir a uno de los sectores anteriores), y Total (suma de los anteriormente citados).
6. **Valor:** Valor absoluto de número de afiliaciones (unidades en miles de afiliaciones) a la Seguridad Social.
7. **Variación interanual:** Valor, en porcentaje, de la tasa de variación interanual del último mes sobre el mismo mes del año anterior.

En la siguiente imagen se muestra una muestra del formato del dataset con las columnas definidas:

Año	Mes	Region	Regimen	Sector	Valor	% Variacion interanual
2021	2	Cataluña	Régimen general y minería del carbón	Total	2.727,10	-2,9
2021	2	Cataluña	Régimen general y minería del carbón	Agricultura	9,5	-0,4
2021	2	Cataluña	Régimen general y minería del carbón	Industria	436,8	-1,9
2021	2	Cataluña	Régimen general y minería del carbón	Construcción	139,8	-1,6
2021	2	Cataluña	Régimen general y minería del carbón	Servicios	2.135,50	-3,2

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

En el siguiente diagrama se ha querido representar el proyecto y dataset elegido. El diagrama se inicia con la adquisición de los datos que componen nuestro dataset, que en este caso corresponden al registro mensual del número de afiliaciones a la seguridad social por sectores y para Cataluña. Estos datos se encuentran disponibles en el sitio web de Idescat y son actualizados mensualmente. Nuestra aportación ha consistido en el desarrollo de un código en Python mediante el cual, empleando técnicas de Web Scrapping, se obtiene un dataset (en formato csv) con todos los datos disponibles y actualizados. El objetivo es emplear estos datos para estudiar posibles variaciones en las tendencias por fenómenos externos como puede ser la situación de pandemia que estamos viviendo.



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

##### 5.1. Campos del dataset

El dataset “Afiliaciones a la Seguridad Social por tipos de régimen y sectores de actividad” está constituido por 7 columnas. En la siguiente tabla se recoge el tipo de datos que contiene cada uno de los campos del dataset:

Nombre del campo	Tipo de dato	Ejemplo formato
Año	Numérico (entero)	2020
Mes	Numérico (entero)	02
Region	Texto	“Cataluña”
Régimen	Texto	“Autónomos”
Sector	Texto	“Construcción”
Valor	Numérico (decimal)	52.6
% Variación interanual	Numérico (decimal)	45.0

Comentario sobre los campos:

- **Año:** Año en formato (YYYY).
- **Mes:** Mes en formato (MM).
- **Región:** Toma los valores: “Catalunya” o “España”

- **Régimen.** Toma los valores: “Régimen general y minería del carbón”, “Autónomos” o “Total”
- **Sector.** Toma los valores: “Total”, “Agricultura”, “Industria”, “Construcción”, “Servicios” y “No clasificables”
- **Valor.** Las unidades son miles de afiliaciones.
- **Variación\_interanual.** Representado en porcentaje (%)

## 5.2. Periodo de tiempo de los datos

Los datos se actualizan de forma mensual en la web de Idescat y en el momento actual se dispone de datos desde 01/1994 hasta 02/2021. La información de las afiliaciones hace referencia al último día de cada mes

**Última actualización:** 10 de marzo de 2021. **Próxima actualización:** 8 de abril de 2021.

## 5.3. Extracción de los datos

Para la extracción de los datos se accede al siguiente sitio web, en el que se encuentran las tablas con los datos. En cada página hay dos tablas: una correspondiente a los datos de Cataluña y una segunda con los datos de España.

<https://www.idescat.cat/indicadors/?id=conj&n=10222&t=202102&lang=es&col=1>

Estas tablas tienen el formato que se muestra en la captura inferior. Cada fila corresponde a un mes (al que corresponden las mediciones) y existen además seis columnas, en las que cada una corresponde a uno de los sectores (Total, Agricultura, Industria, Construcción, Servicios y No Clasificables).



	Total	Agricultura	Industria	Construcción	Servicios	No clasificables
02/2021	2.727,1	9,5	436,8	139,8	2.135,5	5,5
01/2021	2.717,7	9,4	436,0	137,8	2.129,0	5,5
12/2020	2.721,8	9,6	435,6	136,0	2.135,2	5,5
11/2020	2.740,7	9,6	437,5	140,3	2.147,8	5,5
10/2020	2.742,8	9,7	438,5	140,0	2.149,1	5,5

Como se ha comentado en la descripción del dataset, se recogen dos tipos de medidas, por una parte el número absoluto de afiliaciones a la Seguridad Social (“Valor”) y por otra el porcentaje de valor interanual (% Variación interanual). Además, estos datos se recogen para tres regímenes de cotización diferentes (Régimen general, Autónomos y Total). De esta forma existen un total de 6 tablas diferentes para obtener todos los datos correspondientes a estas combinaciones.

En el desplegable situado en la esquina superior izquierda de la captura superior, nos permite seleccionar de qué medición y régimen queremos consultar los resultados. Esto también es fácilmente automatizable empleando la ruta del sitio web y modificando el parámetro col (que toma valores de 1 a 6, uno para cada combinación de parámetros descrita)

<https://www.idescat.cat/indicadors/?id=conj&n=10222&t=202102&lang=es&col=1>

Indicadores de coyuntura económica → Trabajo

**Afiliaciones a la Seguridad Social. Por tipos de régimen y sectores de actividad**

Último periodo Descargar

Afiliaciones a la Seguridad Social. Por tipos de régimen y sectores de actividad  
Cataluña  
Valor. Régimen general y minería del carbón

Valor. Régimen general y minería del carbón  
Valor. Autónomos  
Valor. Total  
% Variación interanual. Régimen general y minería del carbón  
% Variación interanual. Autónomos  
% Variación interanual. Total

	Total	Agricultura	Industria	Construcción	Servicios	No clasificat
02/2021	2.727,1	9,5	436,8	139,8	2.135,5	

Basándonos en esto, hemos implementado nuestro script en Python para realizar web scrapping del citado sitio web de Idescat.

Este script comienza con un bucle for que toma valores de 1 a 6, y concatena este valor al final con la ruta del sitio web para poder obtener cada una de las 6 tablas descritas.

Dentro de este bucle, para cada iteración se emplea la librería *Beautiful Soup*, para encontrar las dos tablas que aparecen en cada página (una con los datos de Cataluña y otra con los datos de España).

Se emplea un nuevo bucle que recorrerá cada una de las tablas encontradas y a continuación, dentro de este, otro bucle que leerá cada una de las filas de las tablas, guardando cada campo en la variable que le corresponda (año, mes, región, régimen, sector, valor y variación interanual).

## 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.

El propietario de los datos es **Idescat**, o lo que es lo mismo, el Instituto de Estadística de Cataluña. Idescat es el organismo oficial de estadística de Cataluña, cuya misión es proveer información estadística relevante y de alta calidad, con independencia profesional, y coordinar el Sistema estadístico de Cataluña, con el objetivo de contribuir a la toma de decisiones, la investigación y la mejora de las políticas pública. Desde su página web pueden obtenerse una gran cantidad de datos principalmente relacionados con el territorio catalán, pero también hay datos referentes a España.

Los datos utilizados para este proyecto han sido recolectados desde su página web (<https://www.idescat.cat/indicadors/?id=conj&n=10222&lang=es>) en la que, según citan, obtienen los datos del Ministerio de Trabajo, Migraciones y Seguridad social.

Siguiendo las buenas prácticas del web scrapping y la licencia por la que se rigen esos datos hemos diseñado un código en Python para obtener los datos que nos interesan bajo demanda. Se ha procurado, además, no saturar el servidor de peticiones intentando minimizar el número de peticiones realizadas, y no alterar los datos ni eludir las condiciones que se especifican en la licencia.

Desde Idescat recomiendan consultar el enlace siguiente, en el que podemos encontrar análisis, estudios y recursos relacionados con el proyecto que hemos escogido.

<https://observatoritreball.gencat.cat/ca/inici/>

A continuación, analizaremos alguno de estos informes.

En el siguiente informe vemos un análisis de los contratos registrados en Catalunya en 2020 desde muchas perspectivas. Si nos centramos en los apartados del análisis por sector, como son datos relacionados con los que nosotros hemos utilizado, podemos comparar las conclusiones a las que llegan con nuestros datos para poder ir viendo la tendencia en este 2021.

<https://observatorit treball.gencat.cat/web/.content/generic/documents/treball/estudis/contractacio/2020/arxiu/Balanc-contractacio-2020.pdf>

**Taula 9**

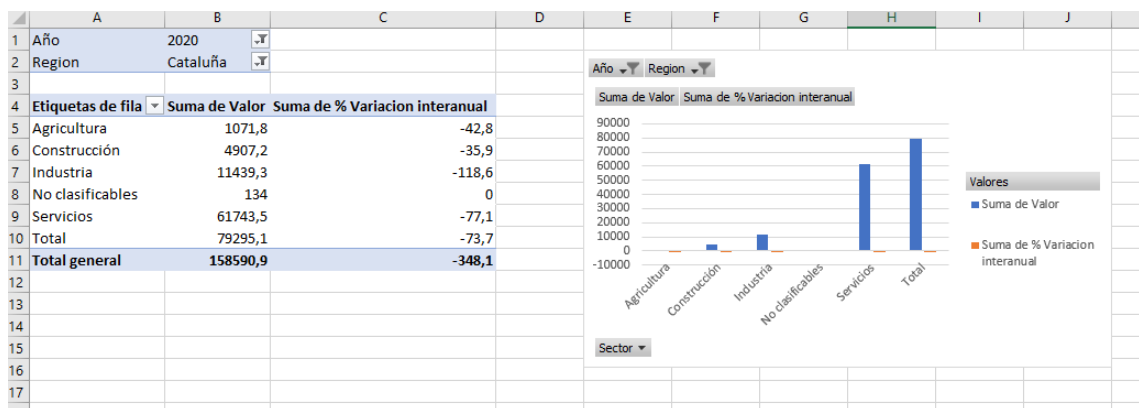
**Catalunya – Contractes registrats per sector d'activitat**

2019 – 2020

	2019		2020		Var. interanual	
	Valor	% distrib.	Valor	% distrib.		
Agricultura	85.141	2,5%	77.330	3,5%	-7.811	-9,2%
Indústria	442.519	13,2%	356.590	15,9%	-85.929	-19,4%
Construcció	141.071	4,2%	118.192	5,3%	-22.879	-16,2%
Serveis	2.677.524	80,0%	1.687.501	75,3%	-990.023	-37,0%
<b>Total contractes</b>	<b>3.346.255</b>	<b>100,0%</b>	<b>2.239.613</b>	<b>100,0%</b>	<b>-1.106.642</b>	<b>-33,1%</b>

Font: Observatori del Treball i Model Productiu a partir de les dades del SOC i del Servei Públic d'Ocupació Estatal

Si contrastamos estos datos con los que obtenemos nosotros en el dataset vemos como, efectivamente, servicios e industria son los sectores de actividad que más despuntan en este caso



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

Una de las grandes preocupaciones de esta pandemia ha sido, es y, esperemos que, por poco tiempo, será su afectación al terreno económico y laboral. Bien es sabido, que la primera consecuencia de la misma es la gran cantidad de pérdidas de puestos de trabajo o de ERTES que se han producido en todo el territorio español. Es por ello que tener unos datos históricos en los que poder ver la evolución, marcar una tendencia e incluso predecir los próximos meses, puede ayudar a ver la luz al final del túnel.

En nuestro caso particular, el motivo de la selección de este dataset se ha debido a que buscábamos obtener datos que nos permitieran valorar objetivamente las repercusiones de la pandemia y al mismo tiempo obtener parámetros que nos permitan monitorizar su mejoría.

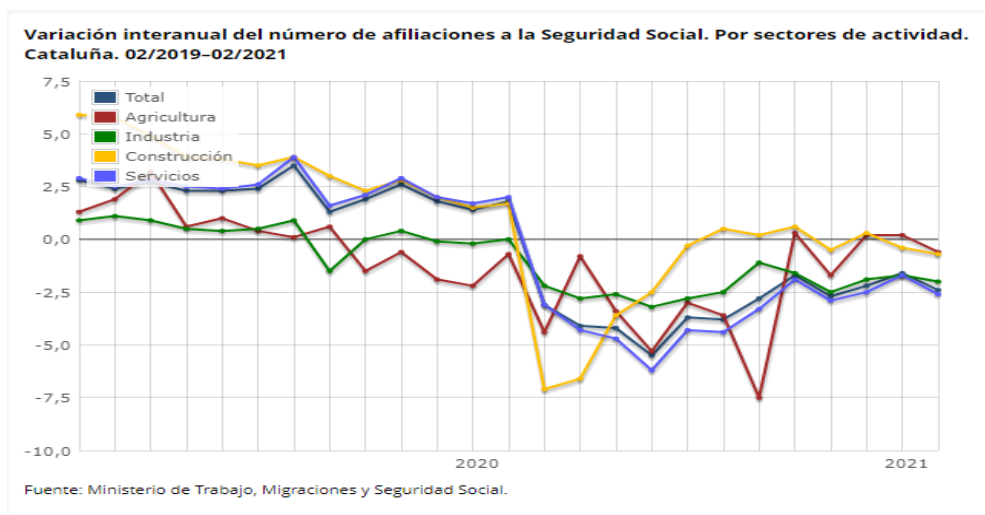
Cuando empezamos a pensar la perspectiva en la que analizaríamos este problema se nos vinieron a la mente dos caminos a seguir. El primero, quizás el más obvio, era obtener datos relacionados con los índices de hospitalización y vacunación. Sin embargo, lo descartamos porque en los últimos meses se ha generado mucha información en esta línea y suponíamos que más gente optaría por este tipo de datos y nosotros buscábamos proporcionar algo

diferente. Entonces empezamos a analizar las diferentes áreas en las que la pandemia COVID ha tenido una mayor repercusión y empezamos a orientarnos hacia ese camino.

El segundo camino se centra en la repercusión económica que pueda tener. Para ello, un buen indicador de en qué medida se van recuperando los puestos de empleo es monitorizar temporalmente el número de afiliaciones a la Seguridad Social y es ahí donde empezamos a buscar datos que poder utilizar para la práctica. Además, ya que la información en este dataset aparece segregada por sectores y régimen, proporciona información muy interesante acerca de qué sectores son los más afectados o los que presentan más dificultades para la recuperación.

Desde el conjunto de datos en csv que generamos se puede analizar la tendencia que se ha seguido en cuestión de altas y bajas de seguridad social desde 1994 y con esos datos se puede realizar una regresión lineal con la que intentar hacer una predicción de cómo va a seguir esa tendencia en los próximos meses.

En la siguiente gráfica, se puede observar la evolución temporal (2019-2020) de la variación interanual del número de afiliaciones a la Seguridad Social en Cataluña, según diferentes los sectores: agricultura, industria, construcción, servicios y el total del conjunto.



## 8. Licencia. Seleccione una de estas licencias para su dataset y explique el motivo de su selección:

Se ha seleccionado la licencia **“Released Under CC BY-NC-SA 4.0 License”** ya que permite que se haga uso de los datos de forma no comercial respetando al creador. De esta forma se ha de reconocer adecuadamente la autoría, proporcionar un enlace a la licencia e indicar si se han realizado cambios.

El motivo de su selección es respetar la política de Idescat, de donde se han tomado los datos que se recogen en el dataset.

Los contenidos y elementos gráficos que constan en el sitio web [www.idescat.cat](http://www.idescat.cat) son de titularidad exclusiva del Idescat o de otros organismos que han autorizado su uso y difusión al Idescat.

Esta web, y todo el material que contiene, está protegida por los derechos de propiedad intelectual e industrial. Por lo tanto, cualquier uso que se haga de sus contenidos está sometido a la normativa reguladora de estos derechos.

El Idescat no permite el uso de logotipos, marcas u otros signos distintivos de esta web y que son propiedad del Idescat, sin que haya tenido conocimiento ni lo haya autorizado previamente.



Con carácter general, el Idescat permite la reutilización de la información que difunde en su sitio web, siempre que se cite la fuente de procedencia, que no se altere ni se desnaturalice el contenido ni el sentido de la información y que se mencione la fecha de la última actualización de la información. Para citar la fuente de procedencia de la información, se puede hacer de la siguiente manera:

- Fuente: Idescat. (Si no hay un tratamiento de los datos.)
- Fuente: elaboración propia a partir de datos del Idescat. (Si los datos se tratan posteriormente.)

El Idescat puede limitar la reutilización de la información por la tutela de otros bienes jurídicos prioritarios, como la protección de los datos personales, la intimidad o los derechos de propiedad intelectual de terceros.

En ningún caso se puede indicar ni sugerir que el Idescat participa, patrocina o apoya la utilización de la información que haga el usuario y sin previo aviso.

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

El código con el que se ha generado el dataset (en Python) con el nombre "web\_scrapper.py" se encuentra accesible en el siguiente repositorio de GitHub:

<https://github.com/taniapvidal/SS-afiliaciones-scrapper>

Para la implementación de este código se han seguido, como se ha comentado en los apartados anteriores, las buenas prácticas recomendadas en Web Scrapping aprendidas en la asignatura.

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

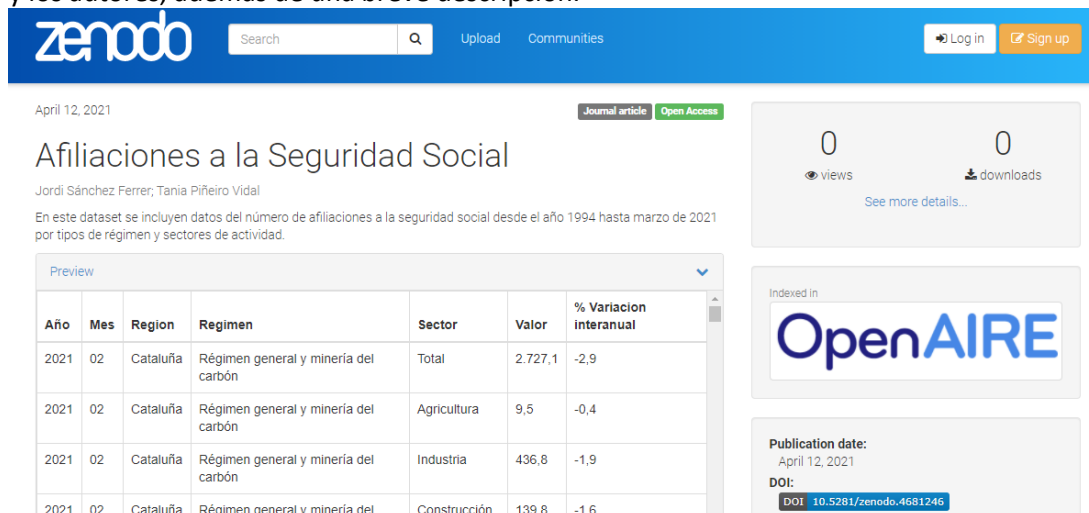
El dataset se ha generado se encuentra accesible con el nombre "Afiliaciones a la seguridad Social.csv" en el siguiente repositorio de GitHub:

<https://github.com/taniapvidal/SS-afiliaciones-scrapper>

Además, se ha publicado en formato csv en Zenodo, al que se puede acceder a través del siguiente enlace:

<https://zenodo.org/record/4681246#.YHRxiOgzaUk>

En la siguiente imagen se muestra una captura de pantalla del dataset publicado en Zenodo, en Zenodo, en la que se puede ver el nombre del mismo "Afiliaciones a la Seguridad Social" y los autores, además de una breve descripción.



The screenshot shows the Zenodo interface for a dataset titled "Afiliaciones a la Seguridad Social" by Jordi Sánchez Ferrer and Tania Piñeiro Vidal. The page includes a search bar, upload button, and login/sign up options. The dataset is dated April 12, 2021, and is marked as a "Journal article" and "Open Access". It has 0 views and 0 downloads. The dataset description states: "En este dataset se incluyen datos del número de afiliaciones a la seguridad social desde el año 1994 hasta marzo de 2021 por tipos de régimen y sectores de actividad." A preview table is shown with the following data:

Año	Mes	Region	Regimen	Sector	Valor	% Variación interanual
2021	02	Cataluña	Régimen general y minería del carbón	Total	2.727,1	-2,9
2021	02	Cataluña	Régimen general y minería del carbón	Agricultura	9,5	-0,4
2021	02	Cataluña	Régimen general y minería del carbón	Industria	436,8	-1,9
2021	02	Cataluña	Régimen general y minería del carbón	Construcción	139,8	-1,6

The page also features an "OpenAIRE" logo and a "Publication date" of April 12, 2021. The DOI is 10.5281/zenodo.4681246.



Finalmente, se ha obtenido el DOI del dataset, que se puede consultar también en ese sitio web y que es el siguiente: **10.5281/zenodo.4681246**

### Contribuciones de los integrantes

Contribuciones	Firma
Investigación previa	TPV, JSF
Redacción de las respuestas	TPV, JSF
Desarrollo código	TPV, JSF