

Oscar Rojo Martín, Álvaro Rodríguez Pardo

zumaia@uoc.edu, alvarorp22@uoc.edu

m2.851 - Tipología y ciclo de vida de los datos

PRA1

Índice

[Puntos para desarrollar 1](#_Toc68540429)

[1. Contexto. Explicar en qué contexto se ha recolectado la información. Explique por qué el sitio web elegido proporciona dicha información. 1](#_Toc68540430)

[1.1. robots.txt 1](#_Toc68540431)

[2. Definir un títulopara el dataset. Elegir un título que sea descriptivo. 2](#_Toc68540432)

[3. Descripcióndel dataset. Desarrollar una descripción breve del conjunto de datos que se ha extraído (es necesario que esta descripción tenga sentido con el título elegido). 2](#_Toc68540433)

[4. Representación gráfica. Presentar una imagen o esquema que identifique el dataset visualmente. 4](#_Toc68540434)

[5. Contenido. Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se ha recogido. 4](#_Toc68540435)

[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. 6](#_Toc68540436)

[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. 6](#_Toc68540437)

[8. Licencia. Seleccione una de estas licencias para su dataset y explique el motivo de su selección: 6](#_Toc68540438)

[9. Código. Adjuntar el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R. 7](#_Toc68540439)

[10. Dataset. Publicación del dataset en formato CSV en Zenodo (obtención del DOI) con una breve descripción. 8](#_Toc68540440)

[Contribuciones al trabajo 8](#_Toc68540441)

Tablas

[Tabla 1: Dataframe con los bots bloqueados por la página web 2](#_Toc68523885)

[Tabla 2: Campos que incluyen los datasets 4](#_Toc68523886)

Ilustraciones

[Ilustración 1: Tabla de subastas desde donde se han recogido los datos 3](#_Toc68523741)

[Ilustración 2: Ejemplo de la información que aparece en los enlaces que contiene la columna “Detalles” 4](#_Toc68523742)

[Ilustración 3: A la izquierda, las subastas celebradas y a la derecha las subastas pendientes 5](#_Toc68523743)

# Puntos para desarrollar

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

Para la realización de web scraping se ha elegido la página de la Diputación Foral de Gipuzkoa (<https://www.gipuzkoa.eus/es>), la cual es es el órgano de gobierno del territorio histórico de Gipuzkoa, País Vasco. La información que presenta este sitio web viene estructurada en forma de categorías definidas en un archivo XML, el cual puede encontrarse en la siguiente página: <https://www.gipuzkoa.eus/sitemap.xml>. Dentro de la página principal, se han escogido dos páginas web que contienen los datos de las subastas de propiedades que están pendientes (<https://www.gipuzkoa.eus/es/web/ogasuna/subastas>) y los datos de las subastas que ya han sido celebradas (<https://www.gipuzkoa.eus/es/web/ogasuna/subastas/celebradas>). Dichos datos se encuentran estructurados en dos tablas, las cuales serán las que se obtendrán a partir del scraper y serán guardadas en un archivo CSV.

* 1. robots.txt

Antes de realizar web scraping sobre la página se ha procedido a estudiar el archivo robots.txt (<https://www.gipuzkoa.eus/es/robots.txt>) para comprobar las limitaciones que podemos tener a la hora de extraer los datos. En este sentido, un archivo robots.txt indica a los rastreadores de los buscadores qué páginas o archivos del sitio se pueden solicitar y cuáles no. Principalmente, se utiliza para evitar que las solicitudes que recibe el sitio lo sobrecarguen; no es un mecanismo para impedir que una página web aparezca en Google. Si lo que se busca es esto último, se debe usar directivas *noindex* o proteger esas páginas con contraseña.

Pasamos a ver los bots que la página tiene bloqueados:

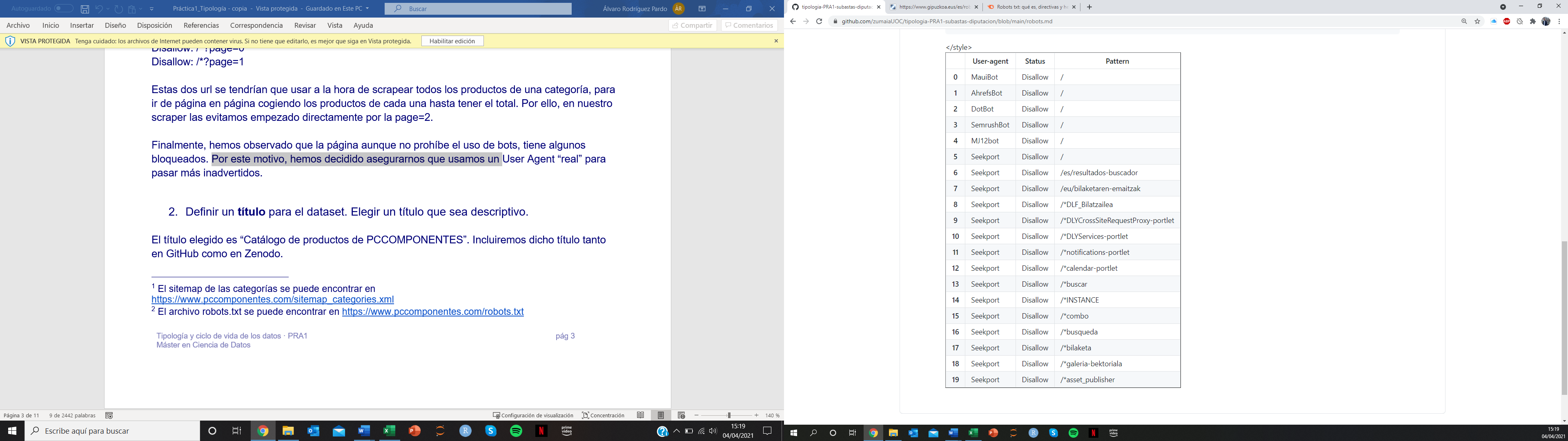


Tabla 1: Dataframe con los bots bloqueados por la página web

Aunque la página no prohíbe el uso de bots, nos hemos asegurado de que usamos un *User Agent* real para pasar más inadvertidos y no tener ningún tipo de problema con estos bots bloqueados.

## Definir un títulopara el dataset. Elegir un título que sea descriptivo.

En este trabajo se han creado dos conjuntos de datos:

* Subastas Pendientes (subastas\_pendientes.csv), donde se encuentra la tabla con las subastas que aún tienen que realizarse.
* Subastas Resueltas (subastas\_resueltas.csv), donde está la tabla que recoge los datos de las subastas que ya han sido celebradas.

Ambos datasets están incluidos en la carpeta *Datasets* del repositorio de GitHub.

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

Como se introdujo en el contexto, los dos conjuntos de datos extraídos hacen referencia a las subastas pendientes y a las ya celebradas por la Diputación Foral de Gipuzkoa, cuyo Departamento de Hacienda y Finanzas realiza subastas de bienes muebles e inmuebles. Ambos datasets contienen una tabla con el mismo número de variables, que son nueve (url, número, tipo, lugar, descripción, importe, fecha, procedimiento y situación), las cuales veremos con más detenimiento en el punto 5. Si uno se fija en la página web, verá cómo la tabla de subastas tiene menos variables de las que se han recogido en el conjunto de datos obtenido, además de tener una llamada “Detalles” en la cual aparece un enlace con información sobre cada una de las subastas. Si se *clicka* en dicho enlace se podrá ver con todo lujo de detalles la subasta, y es precisamente en esta página desde donde se han obtenido variables cómo la descripción, el procedimiento de enajenación y la fecha, esta última estando disponible en la tabla principal, pero viéndose con más detenimiento en la web de detalles.

Como matiz final hay que añadir que, si todas las subastas han sido celebradas, la página web desde donde se puede ver la tabla con las subastas pendientes no tendrá datos, pues todas se han resuelto hasta el momento.

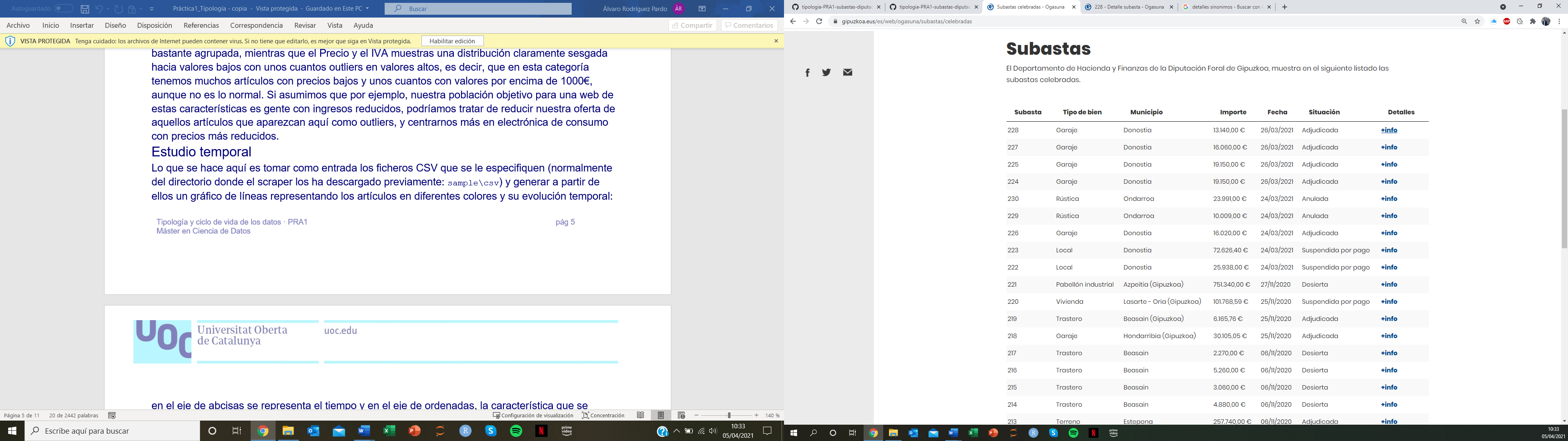


Ilustración 1: Tabla de subastas desde donde se han recogido los datos

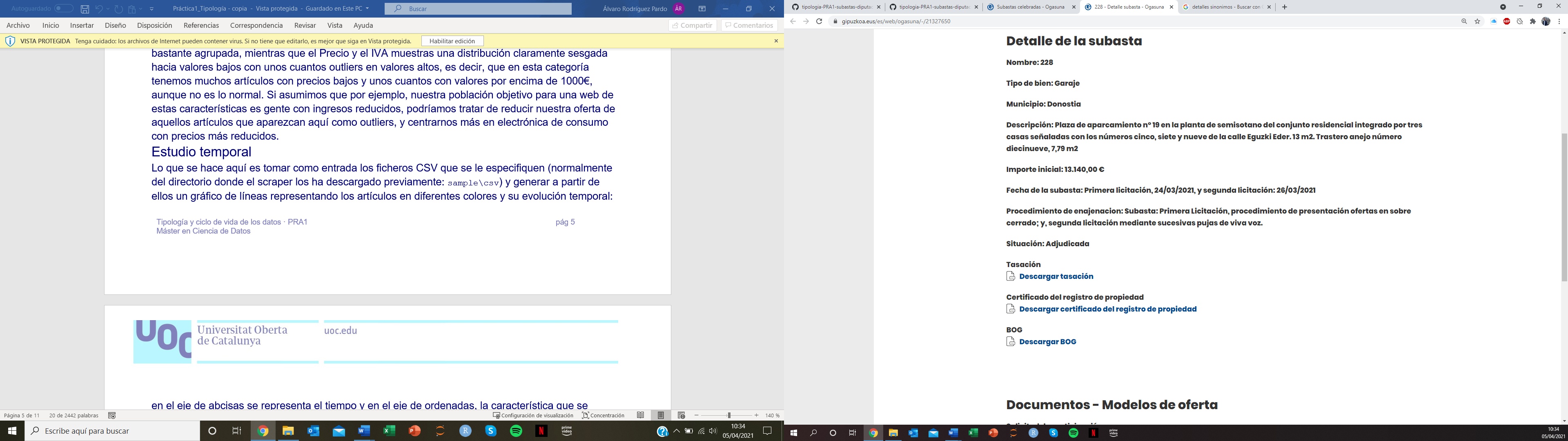


Ilustración 2: Ejemplo de la información que aparece en los enlaces que contiene la columna “Detalles”

## Representación gráfica. Presentar una imagen o esquema que identifique el dataset visualmente.

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

Pasamos a ver una tabla con los campos que tiene cada uno de los dos conjuntos de datos obtenidos:



Tabla 2: Campos que incluyen los datasets

El período de tiempo de los datos de subastas que contiene la página web abarca desde el año 2014 hasta el presente año 2021. Pasamos a continuación a ver cómo se han recogido dichos datos para la obtención de los dos archivos CSV, cuyo procedimiento de recopilación ha sido muy parecido y siguiendo un orden tal como sigue:

1. Se obtiene la URL raíz de la tienda (contendida en un archivo plano).
2. Se sustituye la URL por la palabra tienda.
3. Se recorre la raíz inicial en busca de la URL "hijos".
4. Se recopilan las diferentes URLs en una lista que se ha de limpiar de datos innecesarios y elementos duplicados.
5. Se recorre la lista de hijos en busca de nuevas URLs.
6. Se recopilan las nuevas URLs y se realiza una vez más una limpieza de la nueva lista.
7. De esta última lista de URLs donde se detallan cada uno de los productos, se realiza el scraping.
8. Se genera un diccionario seleccionando los campos que más nos interesan.
9. Por último, se convierte el diccionario en un dataframe y este en un archivo CSV.

En todas las transformaciones, con el fin de preservar la URL y el nombre comercial de la empresa retail, se aplica el punto 2.

Veamos a continuación dos imágenes con los datasets obtenidos en formato CSV:

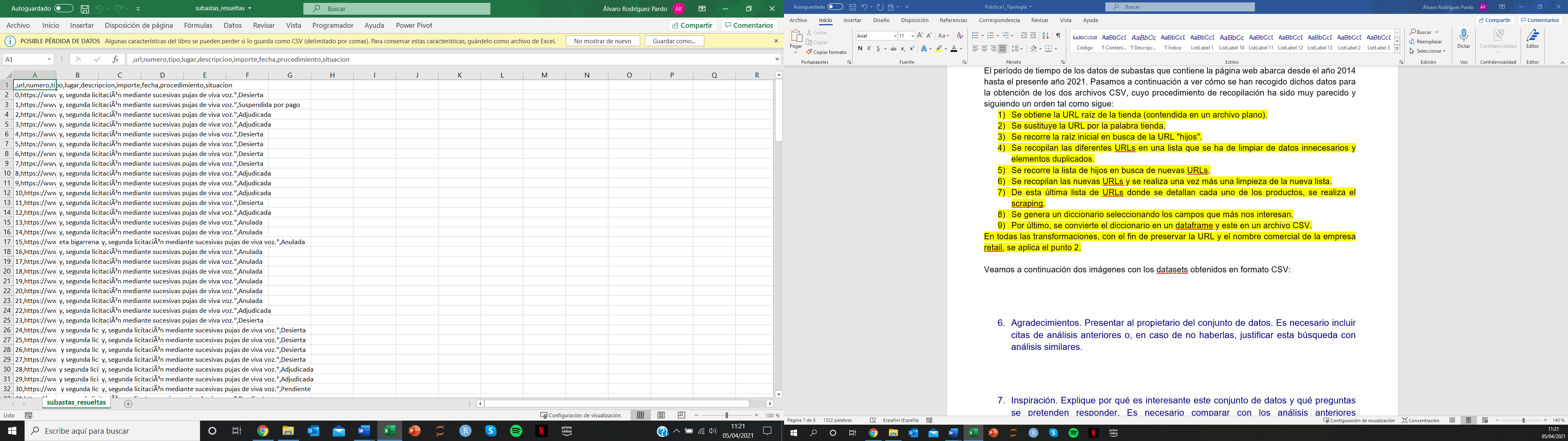
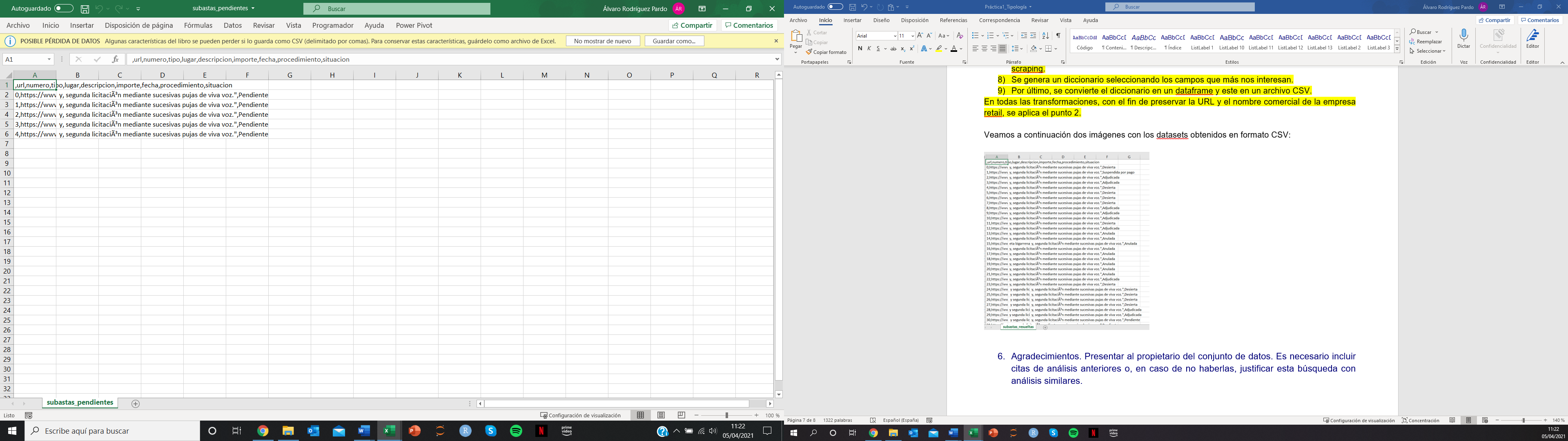
 

Ilustración 3: A la izquierda, las subastas celebradas y a la derecha, las subastas pendientes

## 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 del conjunto de datos es la Diputación Foral de Gipuzkoa (<https://www.gipuzkoa.eus/es/>), concretamente el Departamento de Hacienda y Finanzas (Ogasuna), así que los agradecimientos van dirigidos a esta institución por proporcionar unos datos tan interesantes de forma pública. En cuanto a investigaciones o análisis anteriores, no se ha encontrado ninguno que haga referencia a los datos escogidos.

## 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.

El presente proyecto se inspira en la primera publicación que realizó la Diputación Foral de Gipuzkoa en los portales inmobiliarios de internet como "Idealista" y "Fotocasa", donde se informaba de las subastas. Es un conjunto de datos interesante porque contiene las subastas realizadas en distintos municipios del País Vasco con todo lujo de detalles y podrá responder preguntas del tipo:

* ¿Cuál ha sido el importe máximo y mínimo al que se ha subastado un bien?
* ¿Qué tipo de bien es el más subastado en el País Vasco? ¿Y el menos?
* ¿En qué municipios se realizan más y menos subastas?
* ¿En qué fechas se han realizado más y menos subastas? ¿El por qué de esto se puede explicar por la situación económica de España en esos años? Obviamente, esta última pregunta requeriría de un poco de investigación propia ya que el conjunto de datos en sí mismo no puede explicar la evolución económica del país.

Si observamos el dataset con detenimiento, se puede comprobar que muchas de las subastas que hay en este han quedado desiertas. Habría que valorar si el motivo fue el precio, las condiciones del inmueble o la falta de publicidad del evento.

Finalmente, no se realizarán comparaciones con los análisis anteriores ya que, como se explicó en el apartado 6, no se ha encontrado ningún otro estudio similar.

## 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*

Se ha elegido la licencia CC BY-NC-SA 4.0 ya que esta no permite un uso comercial de la obra original ni de las posibles obras derivadas, pues consideramos que al ser un proyecto de carácter académico no se debe ser utilizado con fines comerciales. Además, la distribución de estas obras derivadas se debe hacer con una licencia igual a la que regula la obra original, con lo cual, nos aseguramos de que se nos dé el crédito por el conjunto de datos obtenido.

Dicha licencia se puede leer detenidamente en el repositorio que contiene el proyecto realizado (<https://github.com/zumaiaUOC/tipologia-PRA1-subastas-diputacion/blob/main/LICENSE.md>)

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

El código para realizar el web scraping se ha escrito en lenguaje Python y se encuentra en la carpeta *Datasets* con el nombre de *diputación.py* (<https://github.com/zumaiaUOC/tipologia-PRA1-subastas-diputacion/blob/main/Code/diputacion.py>). Además, también se presentan otros dos códigos, uno para estudiar el archivo robots.txt, en donde se define una función genérica para obtener un dataframe con aquellos bots que están bloqueados por una página web, el cual también se puede encontrar en la carpeta *Datasets* con el nombre de *utils.py* (<https://github.com/zumaiaUOC/tipologia-PRA1-subastas-diputacion/blob/main/Code/utils.py>), y otro realizado para aplicar dicha función a este proyecto en específico, alojado en la misma carpeta y llamado *robots.py* (<https://github.com/zumaiaUOC/tipologia-PRA1-subastas-diputacion/blob/main/Code/robots.py>).

Para la correcta ejecución del código de web scraping se recomienda lo siguiente:

1. Generar una carpeta:

$ mkdir -directorio

1. Generar un entorno virtual. Cómo hacerlo:

*En Linux:*

$ python3 -m venv /path/to/new/virtual/environment

*En Windows:*

c:\>c:\Python35\python -m venv c:\path\to\myenv

1. Instalar los módulos necesarios detallados en el fichero *requirements.txt*, ubicado en la subcarpeta *data* dentro de la carpeta *Datasets*:

$ pip install requirements.txt

1. Finalmente ejecutar el scraping sobre las subastas de la Diputación:

$ python diputacion.py

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

Se ha publicado el conjunto de datos en formato CSV en Zenodo donde ha añadido una pequeña descripción del proyecto y también el enlace al repositorio de GitHub para quien desee ampliar información sobre cómo se han recogido los datos. El enlace es el siguiente: <https://doi.org/10.5281/zenodo.4662752>. El DOI obtenido se ha añadido al README.md del repositorio.

# Contribuciones al trabajo

|  |  |
| --- | --- |
| **Contribuciones** | **Firma** |
| Investigación previa | ORM, ARP |
| Redacción de las respuestas | ORM, ARP |
| Desarrollo código | ORM, ARP |