M2.851 - Tipología y ciclo de vida de los datos: Práctica 1: Web scraping

Olga Garcés Ciemerozum y Carlos Acosta Quintas

Abril 2021

INTRODUCCIÓN

El presente informe forma parte de la primera práctica de la asignatura M2.851 - Tipología y ciclo de vida de los datos del Máster Universitario en Ciencia de Datos impartido por la Universitat Oberta de Catalunya.

En esta práctica se realizarán técnicas de Web scraping aplicadas a una Web en concreto y se analizarán dichos datos para extraer información relevante y útil.

A su vez, se entregará, junto con la presente memoria, una serie de archivos con el código necesario para la realización de dicho Web scraping y varios juegos de datos reales y actualizados con el que el usuario podrá realizar diferentes estudios analíticos a posteriori.

1 Contexto

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

Contexto en la recolección de la información.

La página web seleccionada para realizar las técnicas de Web scraping es: https://www.bonarea.com/ (https://www.bonarea.com/)

BonÀrea es empresa con una amplia experiencia en el sector agroalimentario, donde su principal negocio se desarrolla en actividades ganaderas, industriales y comerciales con el fin de poder llegar al consumidor sin intermediarios.
Una de las divisiones de la empresa, bonÀrea Energía, dispone de más de 55 gasolineras que venden más de 430 millones de litros al año con un ratio calidad/precio que hace ahorrar más de 40 millones de euros al año a sus clientes.

Las gasolineras bonÀrea son conocidas por su precio económico debido a sus reducidos márgenes de beneficio y al gran volumen de carburantes vendido, y además debido al uso de economías de escala entre sus diferentes líneas de negocio.

Para ésta práctica, se ha decidido hacer un estudio geográfico y temporal de los precios de las gasolineras de bonÀrea Energía, para determinar, no solamente las variaciones diarias de los precios de los diferentes productos, sinó también realizar un registro de las economías de escala asociadas a cada gasolinera (servicios complementarios aportados por cada gasolinera).

De igual forma, y debido a que la página web muestra información descriptiva y relevante de todos los establecimientos que BonÀrea tiene repartido a lo largo del territorio español (supermercados, tiendas, bufets, gasolineras, “Box online” de recogida, Centros de Agricultura y centros Cash para venta al mayor), se ha decidido recolectar y agrupar en un dataset dicha información.

Información extraída del sitio web.

El dominio www.bonarea.com aporta una web dinámica que referencia a todas las divisiones de negocio del grupo, implementando diferentes esquemas de datos para que la experiencia del usuario final a nivel de visualización de la información sea clara y concisa.

Información establecimientos

Tiendas

Presentación
Localizador y horarios
Precios y productos
Redondeo solidario
Solicitud Franquicia
Oferta locales

Bufets-Restaurantes

Nuestros restaurantes
Localizador
Nuestra carta

Gasolineras

Presentación
Sello de calidad
Precio gasóleo domicilio
Pedido gasóleo a domicilio
Localizador y precios

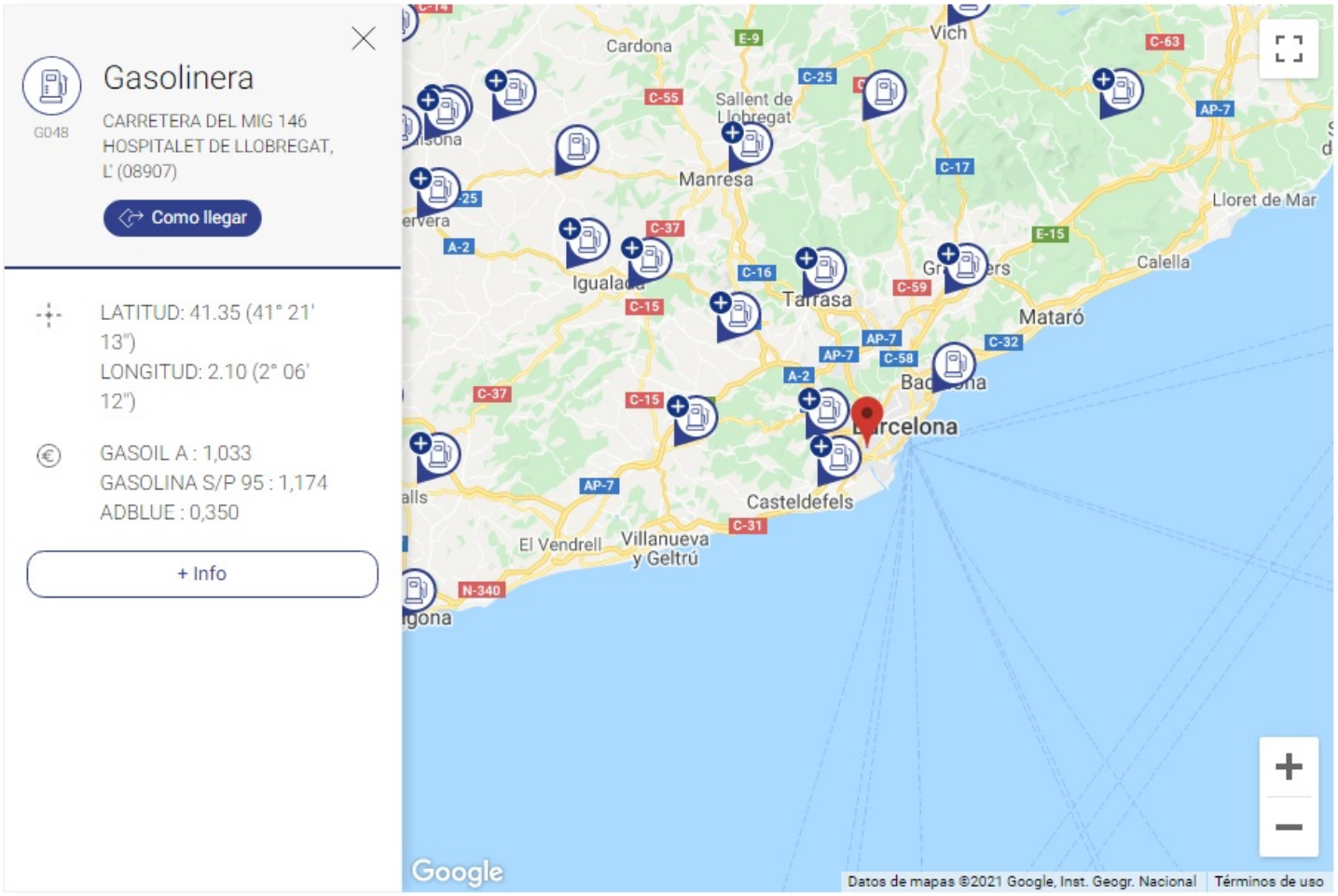
Agrocentros

Presentación
Localizador
Gama de productos
Servicios que ofrecemos
Pedidos pienso

Cash&Carry

Presentación
Localizador
Compra online profesionales

En el apartado de los establecimientos, para todos ellos y en particular en el de las gasolineras, la página web aporta diferentes datos sobre ellas, tanto a nivel geográfico mediante un mapa interactivo para que el usuario sepa en todo momento dónde está la gasolinara más cercana, como a nivel económico (se muestran los precios de los diferentes carburantes disponibles en cada una de ellas).
La página permite al usuario seleccionar uno o varios tipos de establecimientos y muestra la ubicación sobre el mapa de todos establecimientos de los tipos seleccionados.



Además, haciendo click en el icono **"+ info"**, cada establecimiento dispone de una página web anidada para que el usuario pueda tener información adicional como el horario de apertura, dirección , teléfono, etc. de cada uno de ellos.

Tu gasolinera

BARCELONA

HOSPITALET DE LLOBREGAT, L' - CARRETERA DEL MIG 146

DIRECCIÓN GASOLINERA

E.S. BONÀREA L'HOSPITALET DE LLOBREGAT G048
CARRETERA DEL MIG 146
HOSPITALET DE LLOBREGAT, L'

TELÉFONO

973 551 500
Horario: De lunes a viernes de 9h a 13h i de 16h a 19h

PRECIOS

GASOIL A: 1,033
GASOLINA S/P 95: 1,174
ADBLUE: 0,350

HORARIO Y CALENDARIO

ABIERTO TODO EL AÑO
(365 DÍAS Y 24 HORAS)

SERVICIOS QUE OFRECEMOS

Cambio | Lavadero

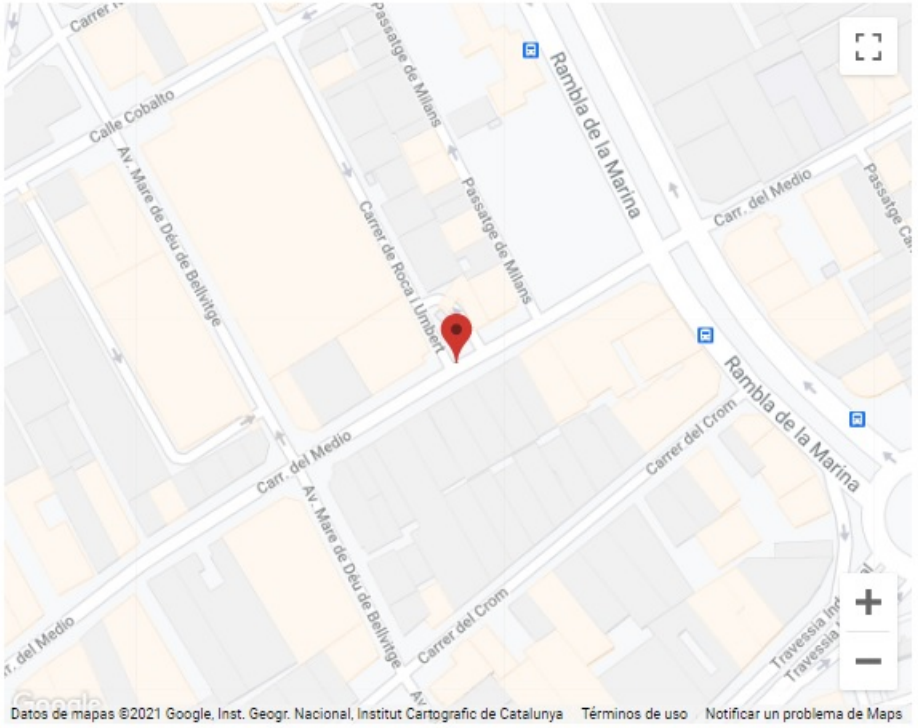
Introduce una población ...

Selecciona los puntos bonÀrea a buscar

Tiendas | Super | Bufets

Gasolineras | Box online | Cash

Agrocentros



En definitiva, el sitio web proporciona dicha información con solamente un motivo: conseguir el compromiso y la fidelización del cliente a una página web práctica y bien diseñada, que ayuda a obtener una información “a la carte” útil.

Hay que tener en cuenta **dos puntos claves** relacionados con la interacción del usuario:

-Durante la interacción del usuario con la web, los datos se generan de forma dinámica: la web envía una solicitud jquery de información y devuelve un conjunto de puntos en el mapa que cumplen con el criterio impuesto por el usuario.

- Al hacer click en los distintos puntos que aparecen en el mapa, la página realiza otra solicitud jquery para una instancia de establecimiento específica y visualiza un recuadro con su información básica.

---

2 **Título del dataset**

---

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

El proyecto puede generar el dataset en dos modalidades:

- **1. Dataset gasolineras en dos archivos csv:** csv con datos de las gasolineras y csv actualizable con histórico de precios diarios de las diferentes gasolineras.
- **2. Dataset completo de todos los establecimientos en un archivo csv** (información descriptiva de los establecimientos)**

Por tanto, habrá 2 títulos propuestos para los diferentes datasets. Los títulos se muestran en orden de aparición en referencia a la anterior lista:

- **1. “Información descriptiva de las gasolineras bonÀrea, servicios asociados y valores diarios del precio de los carburantes”**
- **2. “Establecimientos bonÀrea: Información descriptiva de los supermercados, tiendas, bufets, gasolineras, Box online, Centros de Agricultura y centros Cash.”**

---

3 **Descripción del dataset**

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

“Información descriptiva de las gasolineras bonÀrea, servicios asociados y valores diarios del precio de los carburantes”
El conjunto de datos muestra el identificador de cada gasolinera con la fecha y día de la semana de la extracción de la información.

Además se incluye información básica de cada gasolinera (url, calle, ciudad, código postal, localización geográfica, servicios que provee en el propio establecimiento y los precios diarios de cada producto (Gasoil A, Gasolina sin plomo 95 y 98 y Adblue).

Todos los atributos no tienen valores nulos excepto los servicios asociados y los productos de la gasolinera, donde pueden tomar valores nulos si tal servicio o producto no existe en esa determinada gasolinera.

“Establecimientos bonÀrea: Información descriptiva de los supermercados, tiendas, bufets, gasolineras, Box online, Centros de Agricultura y centros Cash.”

El conjunto de datos muestra la misma información relevante que el anterior juego de datos para todos los tipos de establecimientos a excepción de los precios diarios por tipo de carburante.

4 **Representación gráfica**

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

bonÀrea

Establecimientos BonÀrea

Datos Variables Gasolineras (evolutivos)

Datos Descriptivos (constantes)

Gasolineras

Juego de Datos

- Servicios Asociados
  - ✓ Rentador
  - ✓ Canvi
  - ✓ Super
  - ✓ Lavabo
  - ✓ Parking
  - ✓ Vending
- Precios Diarios Carburantes
  - ✓ Gasoil A
  - ✓ Sin Plomo 95
  - ✓ Sin Plomo 98
  - ✓ ADBLUE

Super

Cash

Box online

Tiendas

Agrocentros

Bufets

Juego de Datos

- Id
- Tipo Establecimiento
- URL
- Calle
- Ciudad
- Código Postal
- Razón Social
- Latitud
- Longitud

5 Contenido

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

El dataset contiene los precios diarios de los carburantes en distintas gasolineras de la red BonÀrea en el período de tiempo desde:

23 de marzo de 2021 hasta la fecha de entrega de la práctica.

Los diferentes datasets pueden ser generados en dos modalidades:

Dos ficheros csv:

'bonarea_gasolineras_prices.csv' - datos históricos del precio de los carburantes por gasolinera

'bonarea_gasolineras.csv' - datos de las gasolineras

Un fichero csv:

'bonarea_gasolineras_data_and_prices.csv' - datos históricos del precio de los carburantes por gasolinera con información detallada de cada gasolinera.

-> **'bonarea_gasolineras_prices.csv'**

- **id** - número identificador de la gasolinera.
- **Fecha** - fecha
- **Dia_Semana** - día de la semana
- **latitude** - latitud

- **longitude** - longitud
- **minutsLatitude** - latitud en minutos
- **minutsLongitude** - longitud en minutos
- **GASOIL A** - precio diario de este tipo de combustible
- **GASOLINA S/P 95** - precio diario de este tipo de combustible
- **GASOLINA S/P 98** - precio diario de este tipo de combustible
- **ADBLUE** - precio diario de este tipo de combustible

-> **‘bonarea_gasolineras.csv’**

- **id** - número identificador de la gasolinera
- **type** - tipo de establecimiento de BonÀrea
- **url** - url de la gasolinera
- **street** - calle donde se ubica la gasolinera
- **city** - ciudad donde se ubica la gasolinera
- **postalCode** - código postal de la dirección de la gasolinera
- **raoSocial** - razón social del establecimiento
- **latitude** - latitud
- **longitude** - longitud
- **minutsLatitude** - latitud en minutos
- **minutsLongitude** - longitud en minutos
- **RENTADO**R - tiene túnel de lavado de coches (1 - sí)
- **CANVI** - tiene cambio para dinero, booleano (1 - sí)
- **SUPER** - tiene supermercado (1 - sí)
- **LAVABO** - tiene lavabo (1 - sí)
- **PARKING** - tiene parking (1 - sí)
- **VENDING** - tiene vending (1 - sí)

-> **‘bonarea_gasolineras_data_and_prices.csv’**

- **id** - número identificador de la gasolinera.
- **Fecha** - fecha**
- **Dia_Semana** - día de la semana**
- **type** - tipo de establecimiento de BonÀrea
- **url** - url de la gasolinera**
- **street** - calle donde se ubica la gasolinera
- **city** - ciudad donde se ubica la gasolinera
- **postalCode** - código postal de la dirección de la gasolinera
- **raoSocial** - razón social del establecimiento
- **latitude** - latitud
- **longitude** - longitud
- **minutsLatitude** - latitud en minutos
- **minutsLongitude** - longitud en minutos
- **RENTADOR** - tiene túnel de lavado de coches (1 - sí)
- **CANVI** - tiene cambio para dinero, booleano (1 - sí)
- **SUPER** - tiene supermercado (1 - sí)
- **LAVABO** - tiene lavabo (1 - sí)
- **PARKING** - tiene parking (1 - sí)
- **VENDING** - tiene vending (1 - sí)
- **GASOIL A** - precio diario de este tipo de combustible**
- **GASOLINA S/P 95** - precio diario de este tipo de combustible
- **GASOLINA S/P 98** - precio diario de este tipo de combustible
- **ADBLUE** - precio diario de este tipo de combustible

Procedimiento de colección de datos

La página permite al usuario seleccionar uno o varios tipos de establecimientos y muestra la ubicación sobre el mapa de todos establecimientos de los tipos seleccionados.

Al hacer click en los distintos puntos que aparecen en el mapa, la página realiza otra solicitud jquery para una instancia de establecimiento específica y visualiza un recuadro con su información básica.

Durante la interacción del usuario con la web, los datos se generan de forma dinámica: la web envía una solicitud jquery de información y devuelve un conjunto de puntos en el mapa que cumplen con el criterio impuesto por el usuario.

Para realizar el webscraping se realiza en dos pasos:

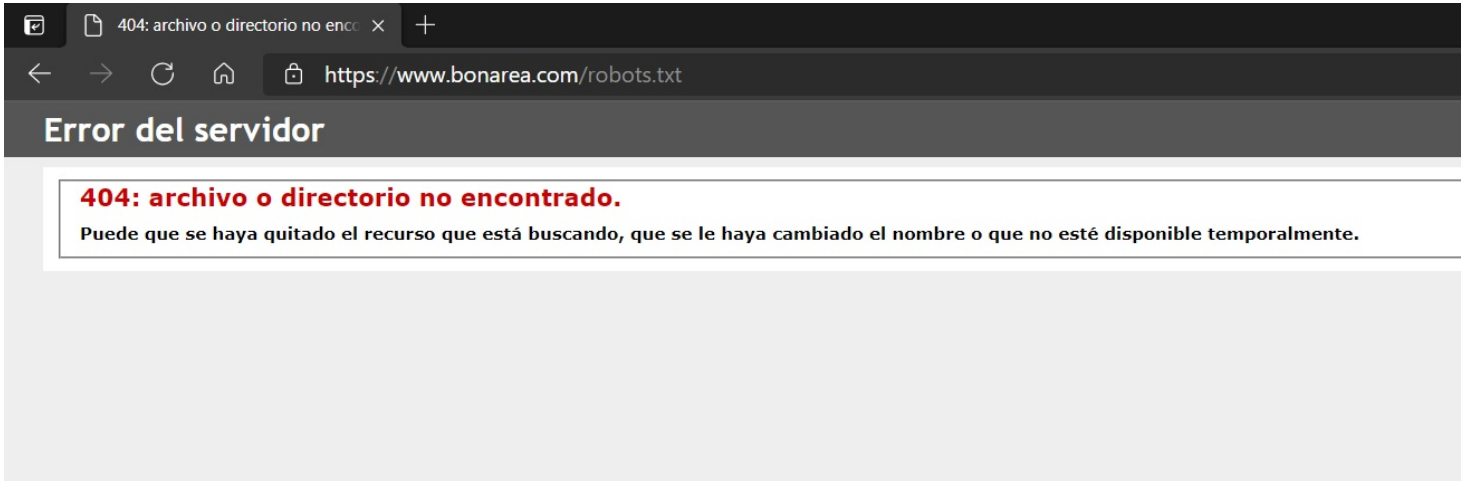
- **Id gasolineras** - encontramos el string que representa la solicitud jquery y usamos Postman (https://www.postman.com/ (https://www.postman.com/)) para obtener los campos `url` , `data` , `header` .
- **Descripción detalle de las gasolineras** - usamos una lista de ids de gasolineras para solicitar información detallada de las gasolineras.

6 Agradecimientos

Los datos han sido recolectados de la página web de BonÁrea. Para ello se han utilizado `requests` de Python y la aplicación `Postman` (https://www.postman.com/ (https://www.postman.com/)).

DESCRIPCION DE LO QUE HACE Y SE CONSIGUE CON POSTMAN

Cabe mencionar que www.bonarea.com no dispone de ningún tipo de archivo “robots.txt”, hecho que ha facilitado el scraping de la información con total libertad por parte de los estudiantes.



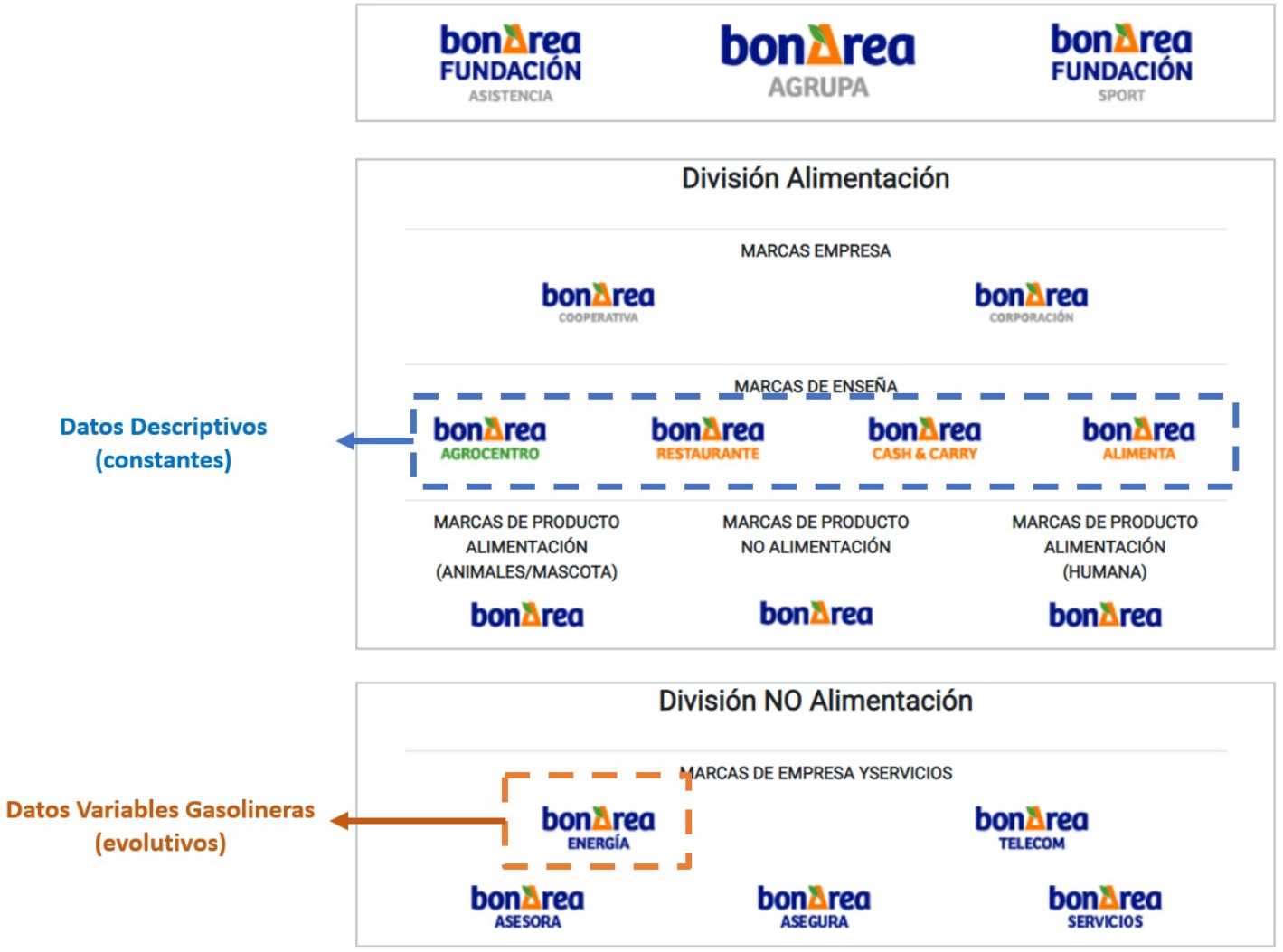
6. Presentar al propietario del conjunto de datos. Es necesario incluir citas de análisis anteriores o, en caso de no haberlas, justificar esta búsqueda con análisis similares.

El fundador y presidente de BonÀrea es Don **Jaume Alsina Calvet**, como tal, el conjunto de datos es de su propiedad.

Si desglosamos los datos según la estructura empresarial de BonÀrea, podemos argumentar que:

-Los datos descriptivos extraídos para los supermercados, tiendas, bufets, Box online, Centros de Agricultura y centros Cash pertenecen a las marcas de enseña (parte de la División de Alimentación); **BonÀrea Agrocentro** (Centros de Agricultura), **BonÀrea Restaurante** (Bufet), **BonÀrea Cash & Carry** (Box Online y Cash) y **BonÀrea Alimenta**(supermercados y tiendas).

-Los datos evolutivos extraídos para las gasolineras pertenencen a la División de NO alimentación, en especial **BonÀrea Energía**.



Después de una búsqueda/rastreo de la página web de BonÀrea y diversas websites relacionadas con la empresa, no existe, al menos públicamente un recopilatorio único **elaborado por BonÀrea** donde se muestre o se pueda descargar toda la información descriptiva de cada establecimiento.

De igual manera, no existe, al menos públicamente, ningún análisis **elaborado por BonÀrea** de la evolución de los precios de sus carburantes por cada una de sus gasolineras, ni por fechas, ni por días de la semana.

Es cierto que algunas websites tratan de recopilar todos los datos posibles relacionados con ciertas gasolineras, pero a veces la información no es completa. Tómese como ejemplo la siguiente página web: (https://www.dieselogasolina.com/ (https://www.dieselogasolina.com/)), que muestra los precios de los carburantes de una gasolinera en concreto del grupo BonÀrea, pero a diferencia de la website de BonÀrea, no muestra todos los carburantes (**el ADBLUE no está analizado**).

Por tanto, la extracción de la información desde página raíz es necesaria para poder extraer un juego de datos completo y 100% fiable y actualizado.

Ejemplo tomado para la gasolinera de GAVA de BonÀrea:

Información extraída de la web de BonÀrea:

Información extraída de la web de Dieselogasolina:

bonArea

¿Datos incorrectos?

Gasolinera **BONAREA**

- Horario: L-D: 24H
- Venta al Público
- Margen de la carretera: Derecho
- Actualiza sus precios **cada 4 días de media**
- Lat,Lng: 41.300000,2.009000

Esta gasolinera no ha notificado precios recientemente.
No podemos asegurarte que estos precios estén en vigor en el día de hoy.

Precios revisados el 05/04/2021.

Última notificación de precios: **31/03/2021.**

¿Te han cobrado o has visto un precio diferente?

SP-95	SP-98	G-A
1,175	1,232	1,024

Otras gasolineras cerca de esta:

REPSOL a 0.52 Km. Ver ficha y precios »
CALLE TARRAGONA, 29

CEPSA a 0.55 Km. Ver ficha y precios »
CARRETERA SANTA CREU DE CALAFELL KM. 5

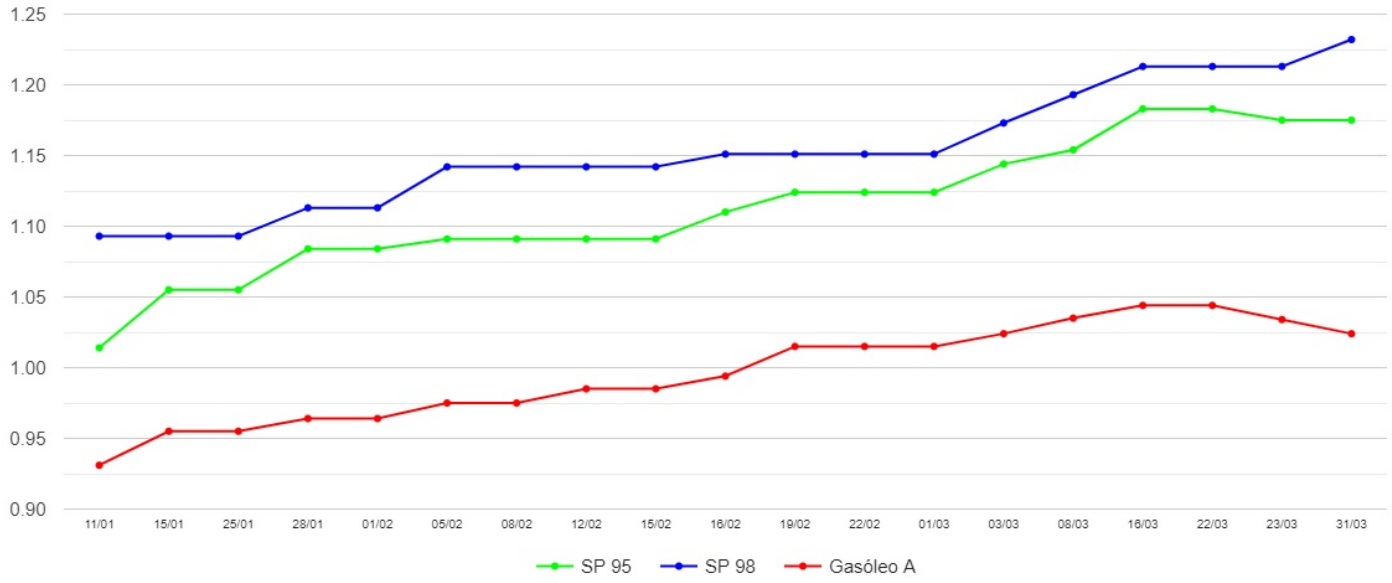
CAMPSA EXPRESS a 0.60 Km. Ver ficha y precios »
AVINGUDA BERTRAN I GÜELL, 46

Q8 a 0.71 Km. Ver ficha y precios »
CARRER ENGINY, 30

CARREFOUR a 1.04 Km. Ver ficha y precios »
AVENIDA DE LA GENERALITAT, 198

Evolución del precio en esta gasolinera

	31/03/2021	23/03/2021	22/03/2021	16/03/2021	08/03/2021	03/03/2021	01/03/2021	22/02/2021	19/02/2021	16/02/2021	15/02/2021	12/02/2021
Sin plomo 95	1,175	1,175	1,183	1,183	1,154	1,144	1,124	1,124	1,124	1,110	1,091	1,091
Sin plomo 98	1,232	1,213	1,213	1,213	1,193	1,173	1,151	1,151	1,151	1,151	1,142	1,142
Gasóleo A	1,024	1,034	1,044	1,044	1,035	1,024	1,015	1,015	1,015	0,994	0,985	0,985



7 **Inspiración**

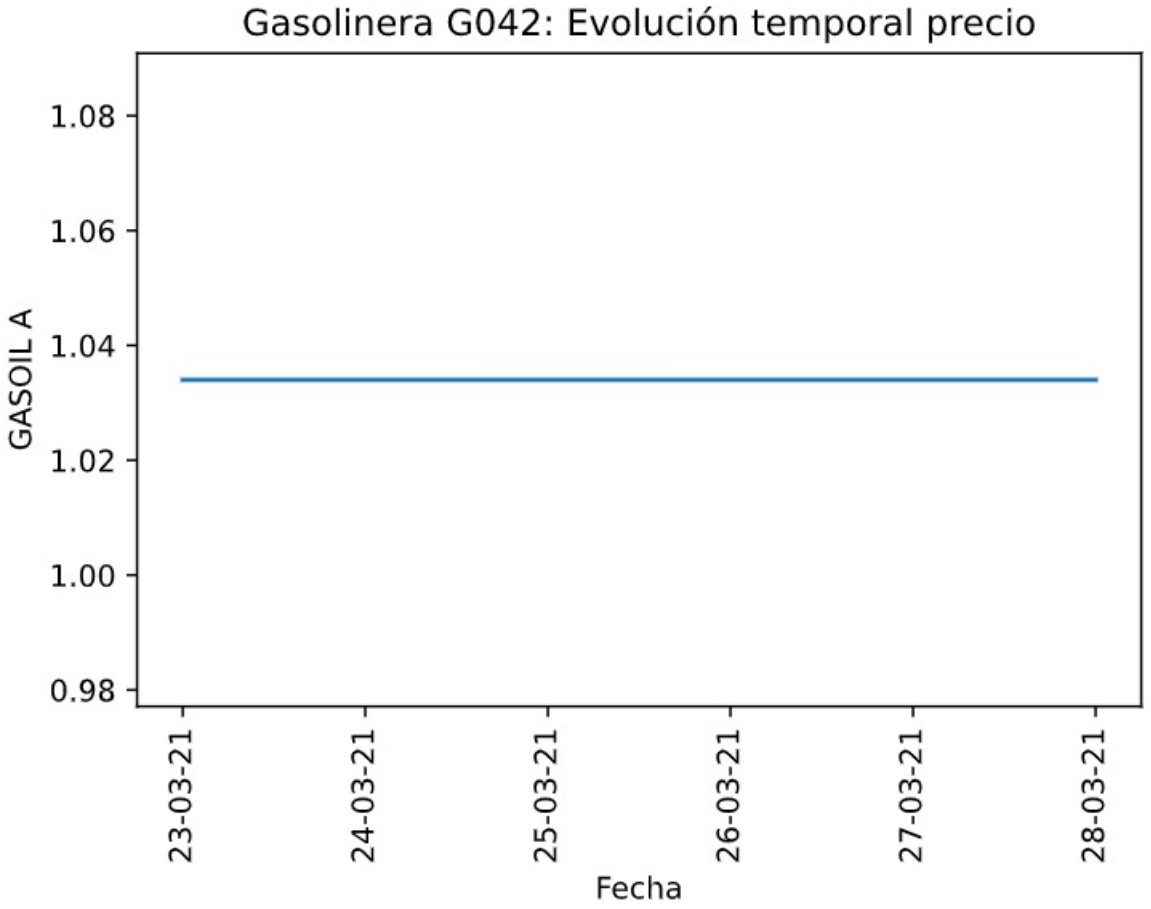
7. Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder. Es necesario comparar con los análisis anteriores presentados en el apartado 6.

El juego de datos generado nos aporta la información sobre los precios diarios de los carburantes (por productos y por gasolineras) de la empresa BonÀrea.
Esta información es realmente interesante para el usuario (y también para la competencia) puesto que se puede generar diferentes aplicaciones o visualizaciones que pueden ser base en toma de decisiones tanto particulares (el usuario puede elegir qué gasolinera repostar por ser más barata) como a nivel estratégico (la empresa competidora puede utilizarlos en su beneficio).

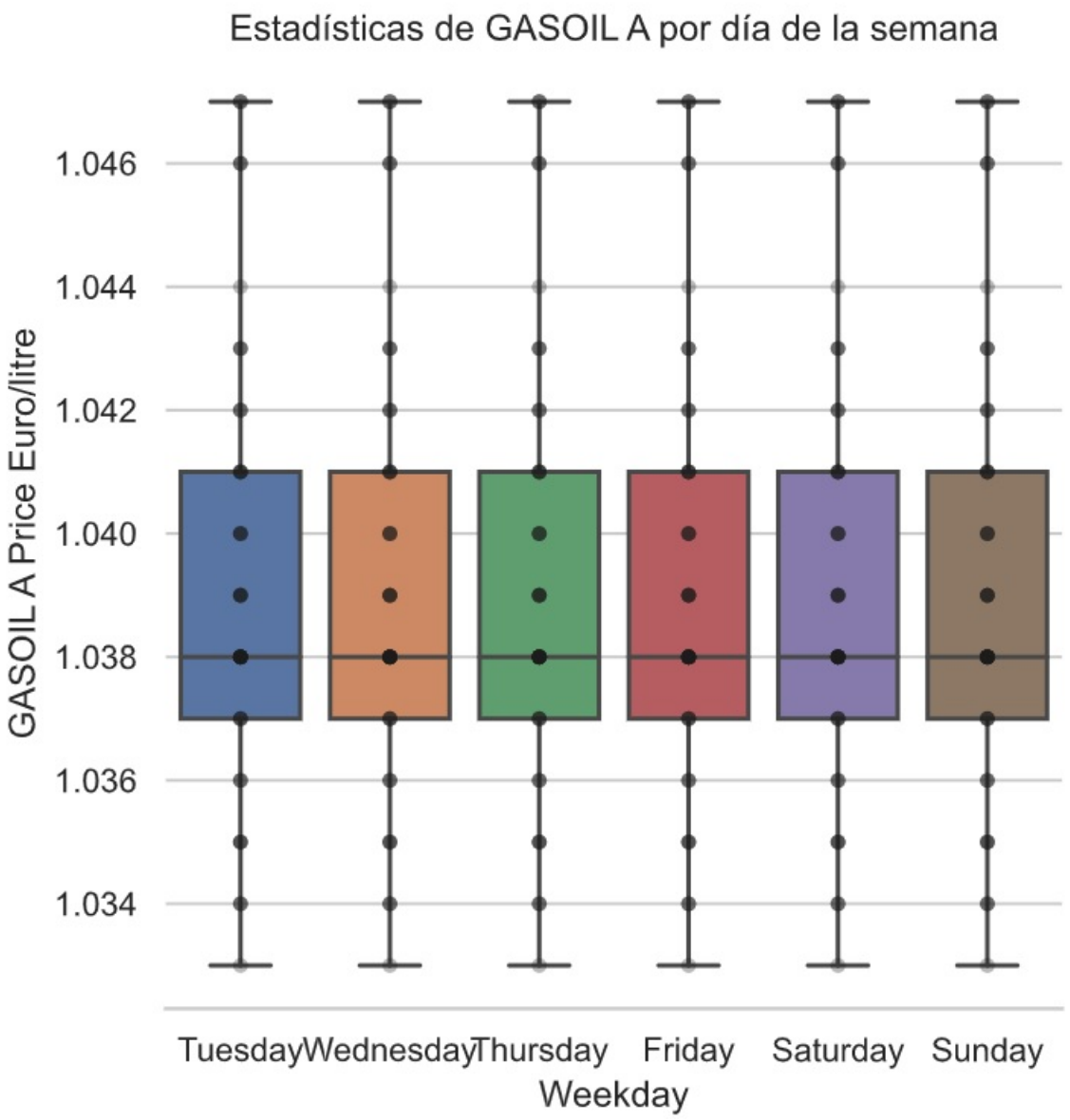
Ejemplos de posibles gráficas que se podría generar, serían las siguientes:

Heatmap geográfico de precios (siendo el color rojo indicador de productos más caros):

Evolución temporal del precio del carburante por producto y por gasolinera:



Estadísticos de los precios de los carburantes por producto y por día de la semana:



Como se puede observar, la información disponible es valiosa y útil en futuras tomas de decisiones.
Por otro lado, el dataset también aporta información de los servicios asociados a cada gasolinera, donde éstos pueden ser:

- **RENTADOR** - Tiene túnel de lavado de coches

- **CANVI** - Tiene cambio para dinero, booleano
- **SUPER** - Tiene supermercado
- **LAVABO** - Tiene lavabo
- **PARKING** - Tiene parking
- **VENDING** - Tiene vending

Y en consecuencia, además de la infomación adicional que el usuario pueda obtener de los diferentes servicios que aporta cada gasolinera, estadísticas sobre los precios de los carburantes mediante la agrupación de los diferentes servicios provistos, puede aportar información oculta, que a primera vista no trivial.

Hay otra aplicación que los estudiantes ven de hecho como un potencial análisis a realizar. Se ha observado que los precios oficiales en la página web no varían todos los días, aunque es sabido que los precios diarios sí que cambian. Los estudiantes se han realizado la pregunta del porqué esta frecuencia de actualizaciones de precios y si pudiera haber algún patrón escondido detras.
Debido a que el juego de datos captura no solamente la fecha, sino también el día de la semana, **y una vez la recolección de los datos pudiese ser grande** (mínimo unos meses de actualizaciones o incluso del órden anual) se podría realizar un algoritmo de clustering para agrupar las diferentes fechas, días de las semanas y gasolineras para poder descubrir un posible patrón de comportamiento.

---

8 Licencia

---

8. 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**

///>>> PENDIENTE DE DECISION.

9 **Código**

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

La generación del dataset se ha generado utilizando el lenguaje de programación Python.

El código utilizado y archivos auxiliares se pueden encontrar en https://github.com/ (https://github.com/) bajo la siguiente dirección web:

https://github.com/Carlos-Acosta/webscraping.git (https://github.com/Carlos-Acosta/webscraping.git)

10 **Dataset**

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

///>>> PENDIENTE DE ELABORACIÓN DE LA ÚLTIMA VERSIÓN DEL DATASET...

11 **Referencias**

-1- Lawson, R. (2015). *Web Scraping with Python*. Packt Publishing Ltd. Chapter 2. Scraping the Data.
-2- Mitchel, R. (2015). *Web Scraping with Python: Collecting Data from the Modern Web*. O'Reilly Media, Inc. Chapter 1. Your First Web Scraper.
-3- Lawson, R. (2015). *Web Scraping with Python*. Packt Publishing Ltd. Chapter 5. Dynamic Data.