Práctica 1 – Tipología y ciclo de vida de los datos

Ramón Serrano Valero

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

He utilizado la web de gasolina barata https://gasolinabarata.info/ ya que es una web que he utilizado con anterioridad para ver precios medios de combustible de poblaciones, por ver en que localidad es mas barata de las que me rodean.

Esta web recopila de forma sencilla, los datos del ministerio de industria, comercio y turismo, por lo que la veía interesante extraer los datos de dicha página.

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

El título del dataset es: "Precio_Medio_Combustible_Municipios_ES".

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 esta formado por los precios medios de cada tipo de combustible fósil, gasolina, diésel, etc… de cada uno de los municipios de cada ciudad española.
En este dataset, encontraremos nombre de provincia, nombre de municipio, fecha de obtención de los datos de cada registro, así como cada uno de los tipos de gasolina con su precio, en caso de disponer de dicho tipo de combustible en el municipio localizado.

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

Icono de la representación gráfica del dataset:

Imagen del dataset:

```
Precio_Medio_Combustible_Municipios_ES
```

```
Creamos Dataframe con los resultados
```

Out[282]:

	City	Town	gasolina 95	gasolina 98	diésel	diésel plus	gasóleo B	biodiésel	GNL	GNC	GLP	Fecha
0	Álava	Alegría-Dulantzi	1,199 €/l	None	1,109 €/l	None	0,726 €/l	None	None	None	None	2020-04-05 18:07:41.365654
1	Álava	Amurrio	1,117 €/l	1,249 €/l	1,044 €/l	1,099 €/l	None	None	None	None	None	2020-04-05 18:07:41.789597
2	Álava	Arraia-Maeztu	1,250 €/l	1,315 €/l	1,145 €/l	None	None	None	None	None	None	2020-04-05 18:07:42.176546
3	Álava	Arrazua-Ubarrundia	1,152 €/l	1,276 €/l	1,084 €/l	1,154 €/l	None	1,159 €/l	None	None	None	2020-04-05 18:07:43.312640
4	Álava	Artziniega	1,165 €/l	1,285 €/l	1,085 €/l	1,125 €/l	0,660 €/l	None	None	None	None	2020-04-05 18:07:43.648186
...	...	...	...	...	...	...	...	...	...	...	...	...
3065	Zaragoza	Villanueva de Gállego	1,111 €/l	1,247 €/l	1,031 €/l	1,147 €/l	0,809 €/l	None	None	None	None	2020-04-05 18:40:52.931560
3066	Zaragoza	Villanueva de Huerva	1,169 €/l	None	1,059 €/l	1,249 €/l	0,679 €/l	None	None	None	None	2020-04-05 18:40:53.341670
3067	Zaragoza	Villarroya de la Sierra	1,079 €/l	1,179 €/l	1,025 €/l	1,095 €/l	0,763 €/l	None	None	None	None	2020-04-05 18:40:53.878471
3068	Zaragoza	Zaragoza	1,107 €/l	1,246 €/l	1,030 €/l	1,118 €/l	0,727 €/l	None	0,869 €/l	0,890 €/l	0,685 €/l	2020-04-05 18:40:54.362681
3069	Zaragoza	Zuera	1,051 €/l	1,239 €/l	0,979 €/l	1,026 €/l	0,562 €/l	None	None	None	None	2020-04-05 18:40:54.859083

3070 rows × 12 columns

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

Las columnas del dataset son:

- City: Provincia española
- Town: Municipio español.
- gasolina 95: Precio €/l del Tipo combustible Gasolina 95
- gasolina 98: Precio €/l del Tipo combustible Gasolina 98
- diésel: Precio €/l del Tipo combustible diésel
- diésel plus: Precio €/l del Tipo combustible diésel plus
- gasóleo B: Precio €/l del Tipo combustible gasóleo B
- biodiesel: Precio €/l del Tipo combustible biodiesel
- GNL: Precio €/l del Tipo combustible GNL
- GNC: Precio €/l del Tipo combustible GNC
- GLP: Precio €/l del Tipo combustible GLP
- Fecha: Fecha de obtención de los datos

Los datos al mostrar la media de precios de tipos de combustible por municipio, aunque la página web los actualice de forma frecuente a diario, nosotros podríamos optar por obtener los datos con una frecuencia de dos veces por semana, ya que no varían tanto los datos de un día respecto a otro.

Por lo que sería interesante automatizar la carga de datos los Lunes y Jueves por ejemplo.

6. Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de investigación o análisis anteriores (si los hay).

Los datos mostrados son obtenidos gracias a la página web gasolinabarata.info mediante técnicas de web scrap. El principal objetivo de analizar los datos es la de realizar la práctica de análisis de datos mediante esta técnica de forma didáctica. Para poder analizar los datos.

Únicamente se han obtenido ciertos datos de la página, como nombres de poblaciones, ciudades, enlaces a sus respectivas páginas y de estas páginas se han obtenido la tabla resumen de precios de combustible.

Analizando el fichero robots.txt, he podido observar que no se permite realizar:
Búsquedas mediante el buscador de la página y lanzar consultas, ni tampoco acceder a ciertas páginas como "feed" o "comments". Tambien he podido observar que hay ciertos bots que están marcados como baneados, como por ejemplo "linko" o "Webcopier", entre otros.

```
Disallow: /cgi-bin
Disallow: /wp-content/plugins/
Disallow: /wp-content/themes/
Disallow: /wp-includes/
Disallow: /*/attachment/
Disallow: /tag/*/page/
Disallow: /tag/*/feed/
Disallow: /page/
Disallow: /comments/
Disallow: /xmlrpc.php
Disallow: /?attachment_id*

User-agent: *
Disallow: /banana/
Disallow: /ir-a/*

User-agent: Googlebot
Disallow: /banana/*
Disallow: /banana/
Disallow: /ir-a/*

User-agent: bingbot
Disallow: /banana/*
Disallow: /banana/
Disallow: /ir-a/*


# Bloqueo de las URL dinamicas
# Disallow: /*?


#Bloqueo de busquedas
User-agent: *
Disallow: /buscar/?q=
Disallow: /buscar/


# Bloqueo de trackbacks
User-agent: *
Disallow: /trackback
Disallow: /*trackback
Disallow: /*trackback*
Disallow: /*/trackback


# Bloqueo de feeds para crawlers
User-agent: *
Allow: /feed/$
Disallow: /feed/
Disallow: /comments/feed/
Disallow: /*/feed/$
Disallow: /*/feed/rss/$
Disallow: /*/trackback/$
Disallow: /*/*/feed/$
Disallow: /*/*/feed/rss/$
Disallow: /*/*/trackback/$
Disallow: /*/*/*/feed/$
Disallow: /*/*/*/feed/rss/$
Disallow: /*/*/*/trackback/$
```

Incluye un fichero de sitemap.xml, donde he podido observar una periodicidad diaria de los datos.

7. Inspiración. Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder.

El conjunto de datos analizado es interesante porque nos permite analizar la evolución histórica del precio de los distintos tipos de combustible, de forma geográfica.
Se pretenden resolver algunas cuestiones como:

- ¿Qué municipio de mi alrededor tiene el tipo de combustible que busco?
- ¿Qué municipio de mi alrededor tiene el precio del combustible más barato?
- ¿Algún municipio tiene variaciones de precio diarias muy desequilibradas con respecto al resto?
- ¿En que municipios debería repostar si planificara un viaje por España?

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

La licencia escogida es la CC BY-NC-ND, licencia que permite descargar el dataset, pero no modificarlo ni distribuirlo comercialmente.
El dataset únicamente contiene los datos de un dia de análisis de web scrap, por lo que se puede copiar y compartir, y debe reconocer la autoria del conjunto de datos al utilizarse.

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

El código está adjuntado en Github, aunque adjunto alguna captura:

```python
#Método que obtiene el listado de ciudades españolas con sus links
def getCities(url,numDivToProcess,textTitle,nameCity):

    dictionaryCities={}

    print("Buscamos en la página:"+url+", los enlaces de las provincias de España")
    page = requests.get(url)

    res = BeautifulSoup(page.content,"html.parser")

    print('En dicha URL, Buscamos los divs que contengan la clase div.listGasolineras')
    divCities=res.find_all("div", class_="listGasolineras")

    print("Hay "+str(len(divCities))+ " divs, hay que coger el "+str(numDivToProcess)+", el de provincias")

    if len(divCities)==0 or "Enlaces internos" in divCities[numDivToProcess-1].text:
        print("Enlace Interno a propia ciudad")
        print(url)
        print(nameCity)
        dictionaryCities[nameCity]=url
    else:
        a_cities=divCities[numDivToProcess-1].find_all("a")
        print("Una vez obtenido el div donde se encuentran los enlaces de las provincias, tenemos que leer cada enlace:

        for a in a_cities:
            if a.has_attr('href'):
                link=a.attrs['href']
                city=a.attrs['title'].replace(textTitle,"")
                print(link)
                print(city)
                dictionaryCities[city]=link
    return dictionaryCities
```

```python
import datetime
#Obtener precios del combustible
def getGasoil(url):
    #Asignamos fecha de obtención del precio
    dictionaryPrices={'Fecha': datetime.datetime.now()}
    try:
        print("Buscamos en la página:"+url+", los precios de combustible de los municipios.")
        page = requests.get(url)
    except HTTPError as e:
        print(e)
    except URLError:
        print("Servidor caido dominio incorrecto")
    else:
        #Obtenemos el contenido de la página
        res = BeautifulSoup(page.content,"html.parser")
        #Si la página retornada es un 404, no pudo encontrar en dicha página el div que contiene precios
        if "The resource requested could not be found on this server!" in res.text:
            print('No se pudo encontrar DIV en '+url)
        else:
            print('En dicha URL, Buscamos el div que contengan la clase div.bloqueGasolina.precioGasolina')

            #Obtenemos div contenedor de precios
            divPrices=res.find("div", class_="bloqueGasolina precioGasolina")

            #Si el div no contiene nada, porque la página no tiene datos, no puede encontrar el div de precios
            if divPrices==None:
                print('No se pudo encontrar DIV en '+url)
            else:
                #Si contiene precios el div, tenemos que trocear los div precios
                print("Una vez obtenido el div donde se encuentran los precios leemos cada precio:")

                for div in divPrices:
                    #Cada div compuesto por dos div contiene un div para el titulo y otro para el precio
                    if(len(div)==2):
                        price=div.find_all('div')
                        #Partimos cada dupla de div en clave / valor
                        dictionaryPrices[price[0].text.replace("Precio ","")]=price[1].text.replace("Precio ","")
                        print(div.find_all('div'))
    return dictionaryPrices
```

10. Dataset. Publicación del dataset en formato CSV en Zenodo con una pequeña descripción.

Publicado en Zenodo:
https://zenodo.org/record/3740645#.XoofkVMzZhE

Práctica 1 – Tipología y ciclo de vida de los datos

Ramón Serrano Valero

```
https://doi.org/10.5281/zenodo.3740645
```

11. Entrega. Presentar el trabajo con el DOI del dataset en Github.

```
10.5281/zenodo.3740645
```

```
https://doi.org/10.5281/zenodo.3740645
```

Contribuciones:

Contribuciones	Firma
Investigación previa	Ramón Serrano
Redacción Respuestas	Ramón Serrano
Desarrollo Código	Ramón Serrano