Contenido

Contexto

El objetivo principal de este proyecto es capturar y almacenar los precios de los combustibles de las estaciones de servicio de toda España diariamente. Los datos no se reconectan con ningun objetivo definido, se pueden usar para dar respuestas a muchas preguntas, pero si está totalmente enfocado en la adquisición de estos mediante tecnicas de web scraping ya que la pagina web no dispone ninguna API para realizar las consultas que deseamos.

Los datos se recolectan a partir del *scraping* a la web "http://clickgasoil.com" que contiene los precios de los distintos carburantes en las todas las estaciones de servicio en España. Esta web dispone de dicha información dado que se dedican a la venta de gasoil calefacción, mediante quienes puedes realizar los pedidos, consiguiendo el mejor precio y el envío a domicilio.

Uno de los motivos principales por los que hemos escogido este sitio web es porque este tiene una política bastante abierta con respecto a los bots, permitiendo su acceso en todos los sus sites. Cierto es que tienen algunas restricciones sobre algunos directorios del site que no permiten el acceso, pero no era la información que nosotros deseamos obtener y tras ver el código de la página web, decidimos que era una buena opción.

Título dataset

El título del dataset, con el objetivo de describir el contenido de este, es “Precios Combustible Spain”.

Descripción

Cada uno de los registros del dataset hace referencia a una estación de servicio en España. Contiene información sobre la marca de la gasolinera, su localización, el precio de los distintos tipos de carburantes y la fecha en la que se registraron.

Figura 1.1: Representación gráfica del dataset

Contenido

Las variables que componen dicho dataset son:

- Var1: MARCA: Contiene el nombre de la marca de la gasolinera correspondiente.
- Var2: Gasolina_95: Contiene el precio de la gasolina 95.
- Var3: Gasolina_98: Contiene el precio de la gasolina 98.
- Var4: Gasoleo_A: Contiene el precio del gasóleo tipo A.
- Var5: Direccion: Contiene la dirección de la gasolinera correspondiente.
- Var6: Población: Contiene la población de la gasolinera correspondiente.
- Var7: Horario: Contiene el horario de apertura de la gasolinera correspondiente.
- Var8: Fecha: Contiene la fecha de la extracción de los datos.

Ya que el objetivo es obtener el precio de diferentes carburantes en todas las estaciones de España para luego poder analizarlos. Se han seguido los siguientes pasos para obtener los datos:

1. Se accede a la pagina web: "http://www.clickgasoil.com/c/precio-gasoil-calefaccion". Aquí podemos encontrar España dividida en comunidades autónomas.
2. Adquirimos el código fuente de la pagina web y buscamos los enlaces de acceso a las paginas webs de las comunidades.
3. Accedemos a cada comunidad y allí buscamos los enlaces de todas las provincias.
4. Accedemos a las paginas web de cada provincia y adquirimos todas urls de los municipios de cada provincia.
5. Accedemos a cada municipio y obtenemos los enlaces de las estaciones que se encuentran en ese municipio.
6. De todos aquellos municipios que tienen datos (cabe decir unos 15 pueblos no tiene datos a pesar de tener pagina web), obtenemos todos los datos que vamos a describir, marca, precios de 3 combustibles habituales, dirección, población y la fecha de la consulta.
7. Los datos que tenemos en listas, los pasamos a un dataframe y este lo exportamos en formato csv.
8. Creamos el archivo de estaciones_servicio.csv para poder acceder a los enlaces de las estaciones de servicio y no tener que volver a obtener los enlaces nuevamente.
9. Creamos el código necesario para que cuando se ejecute el script, este lea los dos archivos csv facilitados("estaciones_servicio.csv" y "Precios_combustible_spain.csv"), de manera que se pueda ir directamente a la captación de nuevos datos y se pueda almacenar en el mismo csv inicial. Este paso se ha realizado por si se quiere siempre obtener en el mismo archivo, pero somos conscientes que estos archivos csv luego se podrían ir almancenando en una base de datos y no requerir que el script tenga que soportar todos los datos siempre. Fácil es la solución de eliminar de la carpeta donde se aloja el archivo para que no lo lea y el script hará el proceso como si fuera la primera vez.

Observaciones:

La práctica ha ido evolucionando, comenzamos ir a por una sola página web, en ella obteníamos las direcciones de todas las estaciones de servicio del área metropolitana de la ciudad de Barcelona, accedíamos a cada una y obteníamos tan solo los precios. Posteriormente decidimos que podíamos obtener las de toda la provincia de Barcelona y lo conseguimos, pero el proceso ya tardaba 1h en realizarse. Posteriormente decidimos que podíamos intentar a realizar España entera, es la entrega que finalmente hacemos y cabe destacar que conseguir los datos de las 10000 estaciones nos ha llevado con una sola máquina, unas 18h.

Es la primera vez que utilizamos el entorno Github y no sabemos si teníamos que dejar todas las versiones que hemos ido haciendo y subiendo, así que las hemos dejado. En la Wiki hemos especificado cuales son los archivos relevantes para la entrega.

Agradecimientos

Los datos se obtienen mediante la técnica *scraping* a la web clickgasoil.com. El objetivo es poder explotar los datos que presenta dicha web con el fin de analizar dicha información.

Para su implementación, se hace uso de la librería BeatifulSoup del lenguaje de programación Python.

Inspiración

Lo interesante de este conjunto de datos resulta de la explotación de los datos desde un punto de vista tanto comparativo como evolutivo entre distintos puntos de venta, los productos que ofrece, así como sus precios.

La explotación de estos datos permite:

- Generar históricos a lo largo del tiempo y conocer en tiempo real quién ofrece los mejores precios o ciertos tipos de carburantes.
  Esto permite optimizar el precio y el servicio (ya sea desde un punto de vista de vendedor o incluso consumidor).
- Conocer el estado de la competencia.
  Es una forma de estudiar el mercado, sabiendo la competencia qué servicios y precios ofrece en cada momento.

Con el análisis de los datos de este dataset, seremos capaces de responder preguntas del tipo:

- ¿Qué gasolinera ofrece un determinado tipo de carburante?
- ¿Qué gasolinera ofrece el precio más/menos competitivo en un determinado tipo de carburante?
- ¿Qué gasolinera ofrece precios más estables/inestables en el tiempo para un determinado tipo de carburante?
- ¿Cuál es el ranking de las gasolineras más/menos competitivas en un determinado momento? (desarrollando algún tipo de ponderación entre precios, servicios y ubicación)

Licencia

Dado que se trata de una dataset simple, con la información de una web concreta en un momento dado, la licencia escogida para este dataset es **CC BY-NC-ND**. Una licencia restrictiva, que sólo permite la descarga, sin la posibilidad de modificar los datos ni utilizarla con fines comerciales.

Las condiciones de dicha licencia se resumen a continuación:

Compartir — copiar y redistribuir el material en cualquier medio o formato

No Comercial — No puede utilizar el material para una finalidad comercial.

Sin Obra Derivada — Si remezcla, transforma o crea a partir del material, no puede difundir el material modificado.

Reconocimiento — Debe reconocer adecuadamente la autoría, proporcionar un enlace a la licencia e indicar si se han realizado cambios. Puede hacerlo de cualquier manera razonable, pero no de una manera que sugiera que tiene el apoyo del licenciador o lo recibe por el uso que hace.

Anexo

Adjunto a este documento, se dispone del archivo donde se presenta el código Python mediante el cual se crea el dataset (distintos formatos: *py*, *ipynb, html)* y el archivo en formato csv ya preparado para su explotación.

A continuación, se presenta brevemente la visualización de los datos:

Desde Python

```
print (df.head(6))
```

```
                Marca Gasolina 95 Gasolina 98 Gasóleo A  \
0           EVOLUTION      1.239€      1.285€    1.157€
1           PETROPRIX      1.229€           -    1.139€
2                GALP      1.374€      1.484€    1.244€
3                GALP      1.394€      1.509€    1.314€
4        SHELL GLORIAS      1.265€      1.369€    1.189€
5  SHELL VILLA OLIMPICA      1.265€      1.369€    1.189€

             Direccion  Poblacion  Horario                       Fecha
0     CALLE PERE IV, 79  BARCELONA   L-V 05 2019-04-11 17:15:17.939432
1    CALLE BADAJOZ, 108  BARCELONA  L-D 24H 2019-04-11 17:15:17.939432
2  CALLE BAC DE RODA, 66  BARCELONA  L-D 24H 2019-04-11 17:15:17.939432
3       Almogavares, 66  BARCELONA  L-D 24H 2019-04-11 17:15:17.939432
4  AVENIDA DIAGONAL, 189  BARCELONA   L-S 07 2019-04-11 17:15:17.939432
5      PASEO CALVELL, 2  BARCELONA   L-S 09 2019-04-11 17:15:17.939432
```

Desde Excel

	A	B	C	D	E	F	G	H	I
1		Marca	Gasolina_95	Gasolina_98	Gasolea_A	Direccion	Poblacion	Horario	Fecha
2	0	EVOLUTION	1.239	1.285	1.157	CALLE PERE IV, 79	BARCELONA	L-V 05	2019-04-11 17:15:17.939432
3	1	PETROPRIX	1.229	-	1.139	CALLE BADAJOZ, 108	BARCELONA	L-D 24H	2019-04-11 17:15:17.939432
4	2	GALP	1.374	1.484	1.244	CALLE BAC DE RODA, 66	BARCELONA	L-D 24H	2019-04-11 17:15:17.939432
5	3	GALP	1.394	1.509	1.314	Almogavares, 66	BARCELONA	L-D 24H	2019-04-11 17:15:17.939432
6	4	SHELL GLORIAS	1.265	1.369	1.189	AVENIDA DIAGONAL, 189	BARCELONA	L-S 07	2019-04-11 17:15:17.939432
7	5	SHELL VILLA OLIMPICA	1.265	1.369	1.189	PASEO CALVELL, 2	BARCELONA	L-S 09	2019-04-11 17:15:17.939432
8	6	SARAS	1.319	1.459	1.229	C/TAULAT, 15 ESQ. CIUDAD DE GRANADA, 2-6-8	BARCELONA	L-D 24H	2019-04-11 17:15:17.939432
9	7	REPSOL	1.409	1.549	1.319	AVENIDA DEL LITORAL, 49	BARCELONA	L-D 24H	2019-04-11 17:15:17.939432
10	8	REPSOL	1.399	1.539	1.309	AVENIDA BOGATELL, 49	BARCELONA	L-V 07	2019-04-11 17:15:17.939432
11	9	REPSOL	1.399	1.539	1.309	CALLE BALMES, 288	BARCELONA	L-D 24H	2019-04-11 17:15:17.939432
12	10	PETROMIRALLES	1.259	1.372	1.259	CALLE PORT DE HAIFA, S/N	BARCELONA	L-D 24H	2019-04-11 17:15:17.939432
13	11	REPSOL	1.399	-	1.319	PASSEIG ZONA FRANCA, 82	BARCELONA	L-D 06	2019-04-11 17:15:17.939432
14	12	B-OIL	1.198	-	1.138	CALLE BAC DE RODA, 130	BARCELONA	L-D 24H	2019-04-11 17:15:17.939432
15	13	GALP	1.324	1.464	1.224	PASSEIG ZONA FRANCA, 81	BARCELONA	L-V 06	2019-04-11 17:15:17.939432
16	14	CEPSA	1.389	1.509	1.299	CALLE COMTE D'URGELL, 219	BARCELONA	L-D 24H	2019-04-11 17:15:17.939432

Contribuciones

El desarrollo de este trabajo se hizo en equipo. Los intergrandes somos Carlos Herrero y Montse Rodríguez, estudiantes del máster Data Science.

Si bien se realizó en la modalidad ‘a distancia’, hemos llevado a cabo distintas videollamadas e intercambio tanto de información como de contenido via whats upp y mail.

Contribuciones	Firma
Investigación previa	Carlos Herrero, Montse Rodriguez
Redacción de las respuestas	Carlos Herrero, Montse Rodriguez
Desarrollo código	Carlos Herrero, Montse Rodriguez