Tipología y ciclo de vida de los datos: PEC 1 – Alberto L. Mariscal Rivas/Elena Naranjo Segura

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

El conjunto de datos recogido es un histórico de los precios de diferentes compañías eléctricas, en función de los diferentes tipos de productos y la potencia ofrecida, obteniendo los precios directamente desde los catálogos disponibles en su página web. Por tanto, parece interesante recoger estos datos y extraerlos para ver la evolución de los precios ofrecidos en el mercado libre, observando si están correlacionados con las subidas y bajadas del mercado eléctrico y si las compañías energéticas repercuten las bajadas de cargos y costes regulados que el Gobierno ha realizado o, si por el contrario, aprovechan esas bajadas para mantener el precio ofrecido y ampliar el margen de beneficios.

Por tanto, desde nuestro punto de vista los datos se han recogido en un contexto de inestabilidad energética y de análisis de un problema de máxima actualidad que nos preocupa a todos. Para ello, se han recogido los datos directamente desde las webs de las principales comercializadoras, en nuestro caso, Iberdrola y Naturgy, en vez de ir a otras fuentes que extraigan los datos de los ofrecidos en los catálogos de las comercializadoras. De esta manera, disminuimos posibles fallos en los datos al ir directamente a la fuente, a la vez que podemos observar las posibles variaciones de precios con el menor retraso posible.

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

Como título del dataset podríamos elegir “Histórico de precios de las principales compañías eléctricas”, el cual refleja el objetivo del mismo (representar un histórico de precios) y de quiénes son los precios que se almacenan (de las principales compañías eléctricas). Sería fácil extrapolar el código y la dinámica para crear un csv similar con los precios de las principales compañías de gas. Debemos destacar que, en nuestro caso, hemos elegido dos de las principales compañías del país pero podríamos incluir en el análisis muchas más.

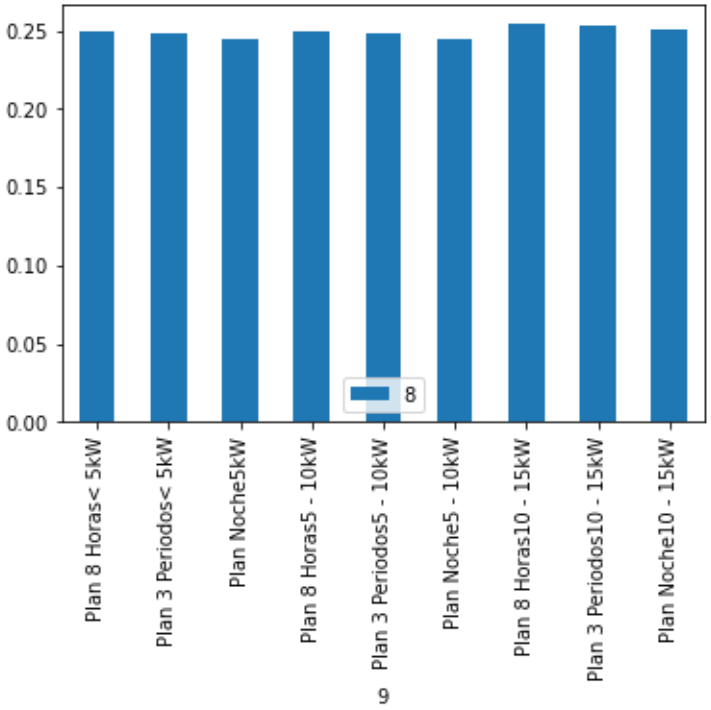
3. Descripción del dataset. Desarrollar una descripción breve del conjunto de datos que se ha extraído. Es necesario que esta descripción tenga sentido con el título elegido

El dataset está estructurado como dataset “longitudinal” que, en contraposición con los datasets horizontales, permite fácilmente guardar valores de precios y ofertas con la fecha, permitiendo acceder a los mismo fácilmente a través de queries de SQL en un futuro y pudiendo convertirlo con “pivot tables” o queries en dataset horizontal. De esta manera, podremos más fácilmente adaptar el código y el dataset a nuevos productos que oferten las

empresas, lo cual sería realmente difícil en un dataset horizontal, en el que tendríamos que ir añadiendo columnas, cosa que resulta mucho más difícil que añadir filas.

4. Representación gráfica. Dibujar un esquema o diagrama que identifique el dataset visualmente y el proyecto elegido

Tal y como se ha comentado antes, se han añadido dos variables de precio medio para poder comparar tarifas, de manera que podamos seguir el histórico de precios de una misma tarifa a lo largo del tiempo o compararla con otra de la misma comercializadora u otra diferente. A continuación se muestra un gráfico de barra para los término de energía y potencia de Iberdrola:



Como se puede ver, es sencillo, visual e intuitivo comparar las diferentes tarifas.

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

Las tarifas eléctricas de pequeños consumidores (catálogo de baja tensión) suelen estar compuestas por un término de energía (el término variable que pagamos por el consumo de la luz) y un término de potencia (el término fijo que deberemos abonar aunque no consumamos). A su vez, estos dos términos pueden estar divididos en varios periodos, habitualmente entre 1 y 3 para el término de energía y 2 para el de potencia. Por tanto, hemos considerado que la estructura óptima del dataset es la siguiente:

- Tarifa: Producto del que se van a almacenar los datos, es un string que incluye información sobre la comercializadora y el nombre del producto, lo cual nos permitirá identificarlo
- Potencia: Los productos habitualmente tendrán precios diferentes en función de la potencia, por lo que deberemos tener otra variable que nos guarde dicha información

- Fecha: Almacenaremos la fecha de la consulta para poder ir alimentando el dataset que en un futuro se almacenará en una tabla en un servidor SQL y al que iremos añadiendo los datos que vayamos obteniendo.
- Te1, Te2 y Te3: Tendremos 3 variables para almacenar el precio de los 3 periodos. En función del producto, puede que cada precio se corresponda con el precio de cada uno de los 3 periodos del calendario o puede, en otros productos, que represente un precio caro y otro barato en función de la oferta de la comercializadora. Podremos incluso encontrarnos con un producto que tienen un único precio, por lo que deberemos tener en cuenta el producto para comprender qué implica cada uno de los 3 términos de energía.
- P1-P2: Precio de los 2 periodos del término fijo
- Media Te: almacenaremos la media del precio de los 3 periodos para poder comparar tarifas y saber cuál es la más cara/bara y además ver fácilmente si una tarifa ha subido o bajado de precio con el tiempo
- Media Tp: igual que con el término variable, almacenaremos la media del término fijo

Los precios se han recogido haciendo uso de Selenium junto con un Driver, lo que permite acceder a las diferentes webs, hacer click en las opciones para seleccionar las diferentes potencias y encontrar el precio fácilmente haciendo uso de la ruta "Xpath" que podremos encontrar al hacer uso de la función "inspeccionar" de Chrome.

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. Justificar qué pasos se han seguido para actuar de acuerdo a los principios éticos y legales en el contexto del proyecto

Queremos agradecer a las comercializadoras por mostrar su catálogo fácilmente en sus webs, haciéndolo accesible y claro, de manera que podamos extraer los datos sin ningún problema. Los datos son de carácter público ya que, obviamente, deben proporcionarlos para que los clientes puedan contratar los precios ofertados con total transparencia. Por tanto, no es necesario seguir ningún paso para actuar de acuerdo a los principios éticos ya que es totalmente aceptable acceder a los precios publicados para su recopilación. Además, como el volumen de consultas es realmente bajo (una vez al día), no supone ningún problema para la página web que pueda colapsarla y hacer que se caiga ya que realmente sería como si un usuario personal consultara los precios una vez al día.

No hemos podido encontrar análisis de la evolución de los precios de las principales comercializadoras del mercado libre por lo que de existir no son fáciles de encontrar. La mayoría de análisis se centran en los precios del mercado regulado o PVPC que son los precios que el Gobierno marca de la casación del pool eléctrico y que no debemos confundir con los que ofrecen compañías como Iberdrola o Endesa que son totalmente libres.

7. Inspiración. Explicar 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 precio de la luz abre telediarios desde hace meses y es un tema de total actualidad, por lo que hemos considerado que es un tema interesante a tratar. Como se ha expuesto antes, existen multitud de gráficas y análisis de la evolución de los precios del mercado regulado pero no hemos encontrado ningún análisis del mercado libre. Con el código y datasets propuestos (que deberán irse completando con los datos de cada día para poder crear un histórico) podremos comprobar si el aumento de los precios del mercado regulado se ve reflejado en los del mercado libre o, si por el contrario, el mercado libre sufre un menor incremento. Este tipo de análisis no los hemos encontrado y consideramos que completan de manera muy interesante todos los artículos periodísticos y noticias que podemos encontrar en el día a día. La electricidad se ha encarecido y nos bombardean con que el PVPC está por las nubes y las eléctricas se están enriqueciendo, pero ¿están las eléctricas subiendo sus precios de igual manera que lo está haciendo el PVPC? ¿Están subiendo menos que el PVPC aunque podrían hacerlo? ¿Están subiendo más incluyendo el margen? ¿Hace el mercado libre y la competencia que la electricidad sea más barata y accesible en épocas en las que el PVPC está disparado?

8. Licencia. Seleccionar una de estas licencias para el dataset resultante y justificar el motivo de su elección

Consideramos adecuada la elección de la licencia “Attribution-ShareAlike 4.0 International” de CreativeCommons (recomendamos ver el documental de Aaron Swartz “The Internet´s Own Boy”). Dicha licencia permite compartir y adaptar el dataset, lo que nos parece interesante porque permite internacionalmente a otras personas añadir datos de otras comercializadoras, nacionales o de otros países, por lo que permitiría comparar la evolución de las principales tarifas españolas con las de otros países Europeos.

Sin embargo, esa licencia obliga a que se nos de crédito porque haber sido los creadores originales del dataset, además de distribuirlo bajo la misma licencia. Creemos que es una dataset idóneo para que la comunidad de internet colabore y cree una gran base de datos.

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

El código se encuentra en el siguiente repositorio: https://github.com/albertolmariscaluoc/IberdrolaPriceScrapper

10. Dataset. Publicar el dataset obtenido en formato CSV en Zenodo con una breve descripción. Obtener y adjuntar el enlace del DOI.

El dataset se ha subido a Zenodo y está disponible en el siguiente link: https://zenodo.org/record/6438904#.YlNSDshBxPY

11. Vídeo. Se debe hacer entrega de un vídeo explicativo de la práctica en donde cada uno de los integrantes del grupo explique con sus propias palabras tanto las respuestas del proyecto como el código utilizado para llevar a cabo la extracción. El vídeo debe ser enviado a través de un enlace a Google Drive que deben proporcionar, junto con el enlace al repositorio Git, al momento de entregar la práctica.

El vídeo se puede consultar en el siguiente enlace:
https://drive.google.com/file/d/1uhfSMNu26aREZzpqyDe5Aibn-76g4vyp/view?usp=sharing