

Tipología y ciclo de vida de los datos

Práctica 1

Máster Universitario en Ciencia de Datos
Universitat Oberta de Catalunya

Jesús González Leal
Francisco Enrique Lorente Banegas

12 de abril de 2020

Pregunta 1

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

Respuesta 1

El contexto de este conjunto de datos se corresponde con los precios de distintas materias primas, y distintos índices de precios, tanto de productos agrícolas como general. Dicha información se ha extraído de distintos sitios web y mediante diferentes técnicas:

- Federación Española de Industriales Fabricantes de Aceite de Oliva
 - Dirección: <http://www.infaoliva.com>
 - Uso de *web scraping*
- Mercados Centrales de Abastecimientos de Alicante, S.A. (Mercalicante)
 - Dirección: <https://www.mercalicante.com>
 - Uso de *web scraping*
- Mercados Centrales de Abastecimientos de Madrid, S.A. (Mercamadrid)
 - Dirección: <https://www.mercamadrid.es>
 - Uso de *web scraping*
- Eurostat (Oficina de estadística de la Unión Europea)
 - Dirección: <https://ec.europa.eu/eurostat/>
 - Uso de API pública

Pregunta 2

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

Respuesta 2

Evolución de los precios de materias primas en España y su influencia en los índices generales de precios.
Price evolution of crop products in Spain and its influence in the general price indices. (para Zenodo).

Pregunta 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).

Respuesta 3

El conjunto de datos generado contiene los precios mensuales de varias materias primas obtenidas de mercados mayoristas y fabricantes, así como precios agregados e índices de precios (generales y agrícolas) con periodicidad igual o menor (trimestral y anual). Algunas de las variables incluyen el mes y año, el nombre del producto y su precio en euros. El rango temporal se sitúa entre los años 2017 y 2019, ambos incluidos, aunque no se dispone de todos los datos en esos tres años.

Pregunta 4

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

Respuesta 4



Figura 1: *Bodegón de frutas y verduras*, de Ginés Andrés de Aguirre (Siglo XVIII).

Pregunta 5

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

Respuesta 5

El dataset contiene datos de los años 2017 a 2019, ambos incluidos. No obstante, no todos los campos disponen de datos en ese rango completo de tiempo, por lo que existen valores vacíos o nulos.

El dataset se compone de los siguientes campos:

- **Producto:** nombre del producto o del índice
 - Formato: Texto
- **Precio:** valor en euros del producto o del índice
 - Formato *Mercalicante y Mercamadrid*: Decimal, separador [,] y símbolo €
 - Formato *Eurostat e Infaoliva*: Decimal, separador [.]
- **Año:** Año del registro
 - Formato: Entero, rango [2017;2019]
- **Mes:** Mes del registro
 - Formato: Entero, rango [1;12]
- **Origen:** Procedencia del registro
 - Formato: Texto, valores posibles: *Eurostat, Infaoliva, MecAlicante, MecaMadrid*

La periodicidad de los datos es mensual. No obstante, en su origen, algunos de los datos presentaban periodicidades distintas, que han sido convertidos a mensuales:

- **Eurostat, IPC anual:** periodicidad anual
- **Eurostat, IPC mensual:** periodicidad mensual
- **Eurostat, Índice de precios agrícolas:** periodicidad trimestral
- **Infaoliva:** periodicidad diaria
- **Mercalicante:** periodicidad mensual
- **Mercamadrid:** periodicidad mensual

Pregunta 6

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

Respuesta 6

Los datos han sido extraídos de los siguientes sitios:

- Federación Española de Industriales Fabricantes de Aceite de Oliva
 - Dirección: <http://www.infaoliva.com>
- Mercados Centrales de Abastecimientos de Alicante, S.A. (Mercalicante)
 - Dirección: <https://www.mercalicante.com>
- Mercados Centrales de Abastecimientos de Madrid, S.A. (Mercamadrid)
 - Dirección: <https://www.mercamadrid.es>
- Eurostat (Oficina de estadística de la Unión Europea)
 - Dirección: <https://ec.europa.eu/eurostat/>

Con respecto al uso de técnicas de *web scraping*, tan sólo la web de Mercamadrid dispone del fichero `robots.txt`, en la cual no prohíbe el uso que hemos realizado. Eurostat también dispone de fichero `robots.txt`, pero hemos accedido a los datos mediante su API pública, por lo que el contenido de dicho fichero no es relevante.

Pregunta 7

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

Respuesta 7

Este conjunto de datos permite por sí mismo identificar patrones de comportamiento del precio de los distintos productos que contiene, así como la relación (o no) entre los cambios en los precios de estos productos y los cambios en los precios agrícolas o generales. Utilizado en combinación con otros conjuntos de datos, podría ayudar a encontrar relaciones entre los precios de los productos y otros eventos, como la meteorología o el precio del combustible.

Pregunta 8

Licencia. Seleccione una de las 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

Respuesta 8

La licencia elegida es **CC BY-SA**, que permite a otros reutilizar y modificar el conjunto de datos, con fines comerciales o no comerciales, siempre y cuando mencionen la autoría y licencien su creación bajo los mismos términos. La intención es favorecer la compartición de datos o conclusiones basadas en ellos. Estas mismas condiciones también se aplican con **Open Database License**, por lo que también se podría utilizarse.

La licencia CC0 no permitiría imponer ninguna condición al uso de los datos, por lo que ha sido descartada. Las demás licencias son más restrictivas, y en este caso, no nos importa que se haga un uso comercial de los datos o que éstos se manipulen.

Pregunta 9

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

Respuesta 9

El código, en lenguaje Python, se encuentra disponible en el siguiente repositorio GitHub:

https://github.com/florenteb/dataset_mat_primas

Pregunta 10

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

Respuesta 10

El dataset está disponible en Zenodo, usando la url siguiente: <https://doi.org/10.5281/zenodo.3749020>.

Pregunta 11

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

Respuesta 11

El DOI del dataset es: [10.5281/zenodo.3749020](https://doi.org/10.5281/zenodo.3749020).

El trabajo está disponible en el repositorio GitHub https://github.com/florenteb/dataset_mat_primas.

Participación en la práctica

Contribuciones	Firma
Investigación previa	Jesús González Leal Francisco Enrique Lorente Banegas
Redacción de las respuestas	Jesús González Leal Francisco Enrique Lorente Banegas
Desarrollo del código	Jesús González Leal Francisco Enrique Lorente Banegas