doc-pract-2

June 5, 2022

1 Tipología y ciclo de vida de los datos: PRÁCTICA II

Componentes de la práctica:

Miguel Rafael Esteban Martín (mestebanmart@uoc.edu)

1.1 Indice de contenidos

- Código
- Descripción del dataset.
- Integración y selección de los datos de interés a analizar.
- Limpieza de los datos.
- Análisis de los datos.
- Resolución del problema: Conclusiones

1.2 Código

La codificación de está práctica se ha realizado a partir de un notebook Jupyter que permite una programación interactiva y documentar y explicar el código mientras este puede ser ejecutado.

Muchas de las gráficas y explicación de porque el uso de cada uno de los métodos usados están en este notebook, es por eso que este documento es una mera guía que en muchos casos hará referencia a secciones de este notebook.

El notebook puede ser facilmente ejecutado desde docker y el repositorio contiene una pequeña documentación para poder ejecutarlo. Por facilidad, también se ha realizado una exportación en un formato PDF de una ejecución de este notebook, una posibilidad que permite Jupyter.

Los enlaces son:

- Repositorio GitHub:
  https://github.com/miguel-esteban-uoc/pract-2/
- Jupyter notebook de esta páctica:
  https://github.com/miguel-esteban-uoc/pract-2/blob/main/notebook/pract-2.ipynb
- Exportación del resultado de la ejecución en formato PDF:
  https://github.com/miguel-esteban-uoc/pract-2/raw/main/doc/pract-2-export.pdf
- Instrucciones de como poder ejecutar el notebook mediante docker:
  https://github.com/miguel-esteban-uoc/pract-2/blob/main/doc/install/docker.md

1.3 Descripcion del dataset

Para esta parte de la práctica se parte del datase creado en la primera práctica.

> ***"Histórico de precios sobre Energía (Gas y Electricidad) para mercados Familiar y Profesional de la Zona Euro"***
> DOI: 10.5281/zenodo.6424152

La razón fundamental de el porqué creamos este dataset es el de entender y comprobar las relaciones entre el precio del gas y la electricidad a lo largo del tiempo dentro de la Zona Euro. Los precios de la energía, es un asunto de actualidad en el tiempo en que se realiza esta práctica, debido al continuo aumento de los mismos. Los datos base fueron obtenidos de la página de la página de estadísitica de la Comunidad Europea, Eurostat: *https://ec.europa.eu/eurostat/web/main/data/database*.

***Nota sobre el dataset en esta práctica*:**

En la primera parte (Jupyter notebook) de la práctica ya se realizó parte del estudio que se pide en el enunciado de esta práctica. El dataset final entregado, incluyó un estudio previo, tanto de los valores extremos, como de limpieza de datos. Se eliminaron los primeros y se propusieron estimadores para los valores nulos.

Como estos apartados entrán dentro del dominio de este trabajo, el enfoque que se ha seguido es, en vez de partir del dataset original, se partirá de los cuatro ficheros de los subdataset que se generaron en pasos anteriores para generar el dataset original: precios del gas doméstico, electricidad doméstico, gas industrial y electricidad industrial.

Los siguientes pasos serán agrupar todos los datos y volver a realizar los estudios de limpieza de datos y gestión de valores extremos. Se obtendrán así, los datos que componen el dataset original, del que filtrando los datos relativos a los precios domésticos, se obtendrá el dataset final de trabajo para realizar el estudio.

1.4 Integración y selección de los datos de interés a analizar

En la primera parte de la práctica ya se hizo un primer estudio de los datos a incluir. Este estudio también se incluye en esta parte de el notebook de esta parte de la práctica (**Sección 2.6** y **2.6.6.**). Aunque se incluyeron los datos de los precios de la electricidad y el gas para un uso no doméstico, no eran suficientes como para poder hacer un estudio, por lo que en esta parte se centrará en los datos de los **precios de la energía (electricidad y gas) en el ámbito doméstico**.

Lo que se pretende es conocer si existe relación entre los precios del gas y la electricidad, en este caso de los valores domésticos. Se ha realizado un estudio de estos precios año por año y también de todo el conjunto de datos.

1.4.1 Estructura del dataset de trabajo

El dataset original tenía una estructura en la que cada precio de un tipo (electricidad ó gas) y año tenía su propia columna y cada registro estaba relacionado con un país:

Campos del dataset original

Nombre del campo	Tipo	Descripción
country	Cadena	El identificador del país
country_name	Cadena	Nombre del país (En inglés)
2017_ElectHouse	Numérico, 4 decimales	Precio de la electricidad doméstica en el 2017
2018_ElectHouse	Numérico, 4 decimales	Precio de la electricidad doméstica en el 2018
2019_ElectHouse	Numérico, 4 decimales	Precio de la electricidad doméstica en el 2019
2020_ElectHouse	Numérico, 4 decimales	Precio de la electricidad doméstica en el 2020
2021_ElectHouse	Numérico, 4 decimales	Precio de la electricidad doméstica en el 2021
2017_ElectNoHouse	Numérico, 4 decimales	Precio de la electricidad industrial en el 2017
2018_ElectNoHouse	Numérico, 4 decimales	Precio de la electricidad industrial en el 2018
2019_ElectNoHouse	Numérico, 4 decimales	Precio de la electricidad industrial en el 2019
2020_ElectNoHouse	Numérico, 4 decimales	Precio de la electricidad industrial en el 2020
2021_ElectNoHouse	Numérico, 4 decimales	Precio de la electricidad industrial en el 2021
2017_GasHouse	Numérico, 4 decimales	Precio del gas doméstico en el 2017
2018_GasHouse	Numérico, 4 decimales	Precio del gas doméstico en el 2018
2019_GasHouse	Numérico, 4 decimales	Precio del gas doméstico en el 2019
2020_GasHouse	Numérico, 4 decimales	Precio del gas doméstico en el 2020
2021_GasHouse	Numérico, 4 decimales	Precio del gas doméstico en el 2021
2017_GasNoHouse	Numérico, 4 decimales	Precio del gas industrial en el 2017
2018_GasNoHouse	Numérico, 4 decimales	Precio del gas industrial en el 2018
2019_GasNoHouse	Numérico, 4 decimales	Precio del gas industrial en el 2019
2020_GasNoHouse	Numérico, 4 decimales	Precio del gas industrial en el 2020

Nombre del campo	Tipo	Descripción
2021_GasNoHouse	Numérico, 4 decimales	Precio del gas industrial en el 2021

Para realizar el estudio es mejor transformar este dataset en uno que tenga en cada registro la información del país, el año de los precios y los precios de la electricidad y el gas, estos últimos son los del contexto doméstico que son en los que se centrará el estudio. Los **campos del dataset de trabajo** son entonces:

Nombre del campo	Tipo	Descripción
country	Cadena	El identificador del país
Year	Entero	Año de los precios
ElectricityPrice	Numérico, 4 decimales	Precio de la electricidad doméstica en ese año
GasPrice	Numérico, 4 decimales	Precio del gas doméstico en ese año

La generación del dataset original se ha realizado en las secciones se ha realizado en las **secciones 1. y 2.** de la práctica y la generación del dataset de trabajo en la **sección 3.**.

1.5 Limpieza de los datos.

Como se ha indicado, en la primera parte de la práctica ya se hizo un análisis de los datos en crudo y se hizo un preproceso de limpieza y de complitud de los datos. Esta parte se ha incluido en este ejercicio la práctica (**sección 2.6.**), mejorando el resultado al poner más antención en los valores extremos de cada uno de los dataset.

Como herramienta de detención de valores extremos se utilizan diagramas de caja. El procedimiento es inicializa a nulo los valores extremos, valores máximos ó mínimos de la población para luego despues en un segundo paso utilizar un estimador de la media para completar los valores que son nulos.

1.6 Análisis de los datos

Para el análisis de los datos primero se crean para los diferentes conjuntos de datos los diagramas de dispersión (**sección 4.1.2.** del notebook de la práctica). Con estos gráficos podemos detectar de una manera visual si existe dependencia y correlación entre los precios del gas y la electricidad de los diferentes conjuntos de datos.

1.6.1 Análisis de la normalidad de las poblaciones

El estudio de la normalidad se ha efectuado en la **sección 4.2.1.** de la práctica. Como primer paso se hace un anális visual generando **gráficos cuartil-cuartil** (**sección 4.2.1.1.**) del notebook

para las poblaciones del gas y la electricidad para los datos de cada uno de los años y el conjunto completo. De este estudio preliminar se puede ver que, en general, los precios del gas se ajustan más a la línea ideal, mientras que para los precios de la electricidad divergen al principio y al final de la gráfica.

El siguiente paso es utilizar el test de **Shapiro-Wilk** (**sección 4.2.1.2.** de la práctica) para comparar la normalidad de manera cuantitativa. Si como resultado del test se tiene que, la población no sigue una distribución normal, se aplicará el **teorema central del límite sección 4.2.1.3.** de la práctica) que dice que una muestra de una población sigue una distribución normal si el número de registros es suficientemente grande, típicamente mayor que 30.

Después de comprobar la normalidad (**sección 4.2.1.4.** de la práctica) cada una de las poblaciones de los diferentes grupos de datos, se llega a la conclusión que todas las poblaciones de los precios del gas, junto a los de la electricidad del año 2020 cumplen el test **Shapiro-Wilk**, algo que concuerda con los gráficos cuartil-cuartil. El resto de poblaciones si que siguen una distribución normal al tener más de 30 registros y cumplir así la condición del teorema central del límite.

1.6.2 Estudio de la homegeneidad de la varianza.

Este evaluación se ha realizado en la **sección 4.2.2.** de la práctica. Se utiliza el test de **Levene** evaluando si las varianzas de los datos del gas y la electricidad para los diferentes dataset es la misma. El resultado es que **ninguno** de los conjuntos de datos **cumplen la condición de homoscedasticidad**.

1.6.3 Aplicación de pruebas estádisticas.

Corresponde en la **sección 4.3.** de la práctica.

Regresión lineal Se hace un estudio de la dependencia lineal entre las dos variables, precios del gas y la electricidad, utilizando un modelo de regresión lineal (**sección 4.3.1.** de la práctica). Para cada conjunto, datos por año y datos completos, se genera el modelo de regresión lineal y se crean las gráficas de la línea generada junto con el diagrama de dispersión calculando el valor de R^2, que es el indicador de ajuste del modelo.

El resultado es que cada uno de los modelos generado tienen un valor de R^2 **pequeño**, el mayor de `0,284647`, correspondiente al año 2021 y por lo tanto todos los modelos tienen un **ajuste pobre**. Es algo que ya se podía intuir a la vista del anális visial anterior de los gráficos de dispersión generados.

Estudio de la correlación En análisis de la correlación se hace en la **sección 4.3.2.** de la práctica.

Todos los conjuntos de datos siguen una distribución normal pero no cumplen la condición de homoscedasticidad. Para comprobar si existe correlación entre las poblaciones de los dos precios se debe usar un test no paramétrico como el test de **Spearman** (**sección 4.3.2.1.** de la práctica). La única condición que exige, es que las distribuciones puedan ordenarse (tiene que existir una relación ordinal) y los precios, al ser númericos, la cumplen.

El resultado es que existen conjuntos de datos, como los de los años 2020 y 2021 y también el conjunto completo, en que los valores están correlados. Destaca el coeficiente de correlación del año 2021 con un valor significativo de `0,6462`, mayor que el resto `0,3841` y `0.3406`, coeficientes

del año 2020 y dataset general respectivamente. En los conjuntos de los demás años, 2017, 2018 y 2019, no existe correlación entre ambos precios.

Contraste de hipótesis Compete a la **sección 4.3.3.** de la práctica. Se estudiará, utilizando contraste de hipétesis si hay o no diferencias significativas a nivel estádistico entre las poblaciones de los precios del gas y la electricidad. Se debe de tener en cuenta que hasta ahora tenemos conjuntos de datos que siguen la normalidad pero que presentan heterocedasticidad, algunos de ellos las variables son dependientes, mientras que otros son independientes.

En estas circustancias, cuando las variables son **independientes** se utiliza el test de **Mann-Whitney** (**sección 4.3.3.1.** de la práctica), que se aplicará para los años que cumplan estas condiciones: 2017, 2018 y 2019. Cuando las variables son **dependientes** se ejecuta el test de **Test de Wilcoxon** (**sección 4.3.3.2.** de la práctica), y que se usará para los dataset completo y de los años 2020 y 2021.

El resultado de ejecución de ambos test es en todos los casos podemos llegar a la conclusión de que **hay diferencias significativas** entre las poblaciones del gas y la electricidad, es decir, que estadísticamente no podemos decir que son la misma población.

1.7 Resolución del problema: Conclusiones

Las conclusiones que podemos sacar de los resultados obtenidos es que todas las poblaciones tienden a una distribución normal pero, sin embargo, entre las poblaciones de ambos precios no existe homogeneidad en las varianzas. Los diagramas de dispersión muestran una distribución uniforme y los modelos de regresión lineal dan un valor R^2 muy pobre, lo que quiere decir en que no hay una dependencia lineal entre las dos poblaciones de precios.

Con respecto a la correlación los precios del año 2020 y 2021 y el dataset completo muestran una correlación con coeficientes de correlación pequeños, excepto el del año 2021. Por último el contraste de hipótesis utilizado para demostrar que ambas distribuciones son diferentes lo ha confirmado en todos los conjuntos de datos.

Como conclusión final existe una pequeña correlación entre los precios del gas y la electricidad domésticos pero no tan fuerte como en un principio se pudiera creer.