Práctica Final de Visualización de Datos

Parte 1: Selección del conjunto de datos

#### Ángel Moreno Prieto

En esta primera parte de la práctica, presentaremos el juego de datos que usaremos para generar la visualización que realizaremos durante la segunda parte.

### Justificación del conjunto de datos seleccionado

Para realizar esta práctica, se han escogido dos conjuntos de datos de temática común: datos globales e históricos sobre la situación de los países frente a la generación, demanda y uso de energía eléctrica. En concreto:

* ***Global Energy Generation & Capacity (IMF)***

Desarrolla datos globales e históricos sobre la generación y capacidad de generación de diferentes países a nivel mundial, y categoriza los resultados en función el tipo de tecnología utilizada.

<https://www.kaggle.com/datasets/pinuto/global-energy-generation-and-capacity-imf>

* ***Global Electricity Demand and Generation Dataset***

Contiene datos globales e históricos sobre la generación y demanda de energía eléctrica de diferentes países a nivel mundial.

<https://www.kaggle.com/datasets/shahriarkabir/global-electricity-demand-and-generation-dataset>

El motivo de estudiar estos datos es que son muy relevantes hoy en día: la electricidad es crítica en nuestras sociedades contemporáneas, y es un requisito básico para el progreso de cualquier nación; pero presenta varios y muy diversos problemas, desde geopolíticos (como la crisis energética en Europa derivada de la guerra ruso-ucraniana de 2021 [1]) hasta medioambientales (el cambio climático obliga y requiere de nuevas formas de obtener energía “limpia” [2]). Analizar cómo diferentes países están afrontando esta situación, y cómo van evolucionando a este respecto a lo largo de los años, puede ser útil para planificar el futuro energético del que tanto dependeremos.

### Relevancia de los datos

Los conjuntos de datos utilizados mantienen información desde comienzos de siglo hasta 2023, por lo que podemos considerar que son bastante recientes.

Si bien no tratan un tema exclusivo para ciertos colectivos, sí que analiza una situación que nos afecta a todos (aunque no a todos *por igual*, pues hay países tanto menos demandantes de energía, como menos capaces de generarla, y, en algunos casos, también incapaces de generarla limpiamente).

### Complejidad

Los dos conjuntos de datos acumulan información sobre más de 200 países, a lo largo de más de 20 años, por lo que tenemos, de forma estimada (salvando algunos valores nulos que tendremos que corregir), unos 4000 conjuntos. A su vez, para cada conjunto, mediremos tres variables:

* **Energía demandada**, en gigavatios hora (GWh). Representará la energía demandada por un país en un año.
* **Energía generada**, en gigavatios hora (GWh). Representará la energía generada en territorio de dicho país en un año. Se clasificará según su modo de generación (gas, carbón, eólica…), por lo que puede haber varias entradas por país/año, si hay varios modos.
* **Capacidad eléctrica instalada**, en megavatios. Representará la cantidad máxima de energía que podía producirse en un país durante un año. Se clasificará también según su modo de generación, por lo que podrá haber varias entradas por país/año.

Como podemos observar, dispondremos de un total de, al menos 12000 registros, con hasta 10 variables (en función de cómo organicemos el dataset) que combinan categóricas (país, continente, modo de generación de electricidad) y numéricas (energía demandada, generada y capacidad instalada).

### Originalidad

El dataset sobre el que trabajaremos será una mezcla de los dos conjuntos de datos anteriormente mencionados, creando la unión en los pares *(país[[1]](#footnote-1), año)*, por lo que será original desde el punto de partida. Por otro lado, añadiremos la variable *“continente”*, con el fin de realizar análisis a una escala más global. Finalmente, también serán necesarias ciertas reconversiones de datos (pasar de TWh a GWh, por ejemplo), y lidiar con los valores nulos que aparecen en algunos países, algunos años. El dataset resultante no deberá tener, de cara a realizar visualizaciones a partir de sus datos, valores vacíos.

### Respuestas objetivo

El conjunto objetivo tendrá la siguiente forma matricial, con los países y el tipo de dato en el eje de las Y, y la fecha, en el eje de las X:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **País** | **Código de país** | **Región** | **Tipo de dato** | **Tipo de energía** | **T2000** | **…** | **T2023** |
| Albania | ALB | Europe | Electricity Generation | Solar | 17606.311 | … | 62146.697 |
| Albania | ALB | Europe | Electricity Demand | - | 18901.44 | … | 55321.07 |
| Argentina | ARG | South America | Electricity Installed Capacity | Fossil Fuel | 13611.947 | … | 26444.228 |
| … | … | … | … | … | … | … | … |
| *Descriptores* | | | | | *Valores* | | |

Donde tendremos 5 variables descriptoras:

* **País** y **Código de país**, representarán al país y su código según la ISO-3166 de tres caracteres.
* La **región**, a la que también nos referimos como **continente**, indicará el superconjunto al que pertenece dicho país. Trabajaremos con seis continentes: Europa, África, Asia, Oceanía, América del Norte, y América Central y del Sur.
* El **Tipo de dato** nos indicará si el valor se refiere a generación, demanda, o capacidad instalada de producción de energía eléctrica.
* El **Tipo de energía** nos indicará la tecnología utilizada para generar la energía, o la asociada a la capacidad indicada. Nótese que la demanda no tiene tipo de energía, pues es un valor a priori (independiente, por tanto, de *cómo* se produzca esa energía).

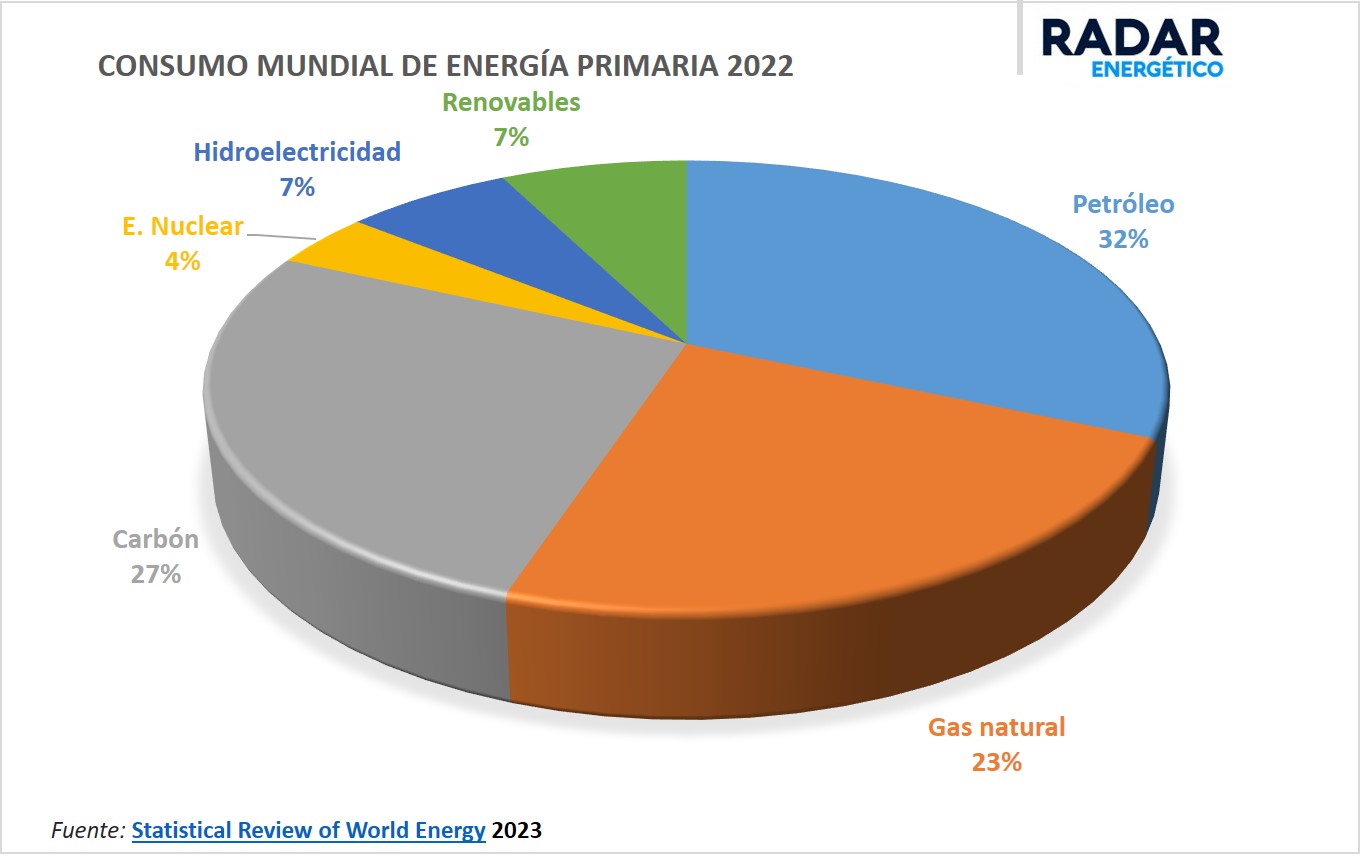
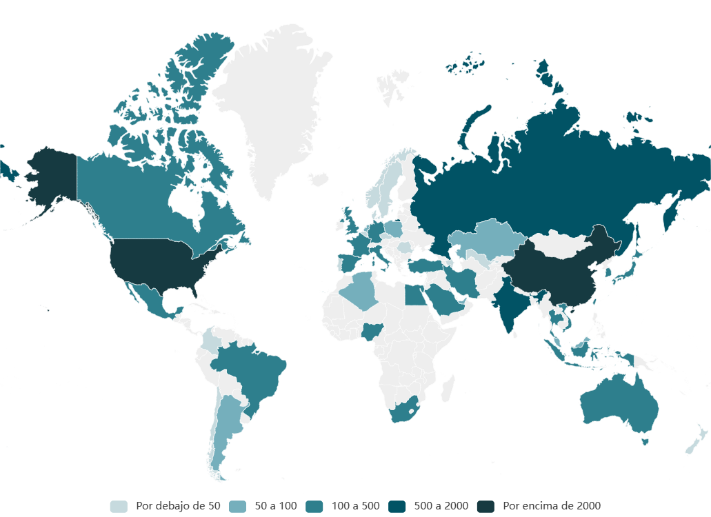
Y 24 variables objetivo, una para cada año, entre el 2000 y el 2023.

Por otro lado, dado que dispondremos de un conjunto de datos tanto históricos como geográficos sobre un tema de interés actual, nos plantearemos las siguientes preguntas:

Desde el punto de vista categórico / **geográfico**:

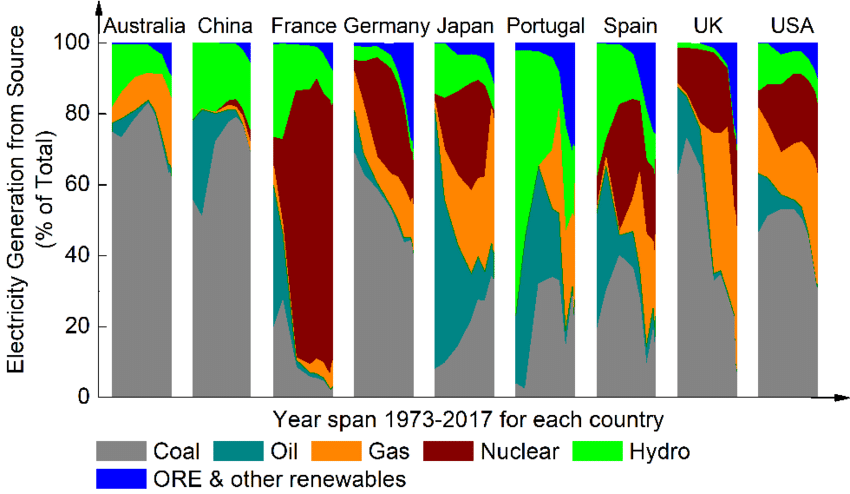
* *¿Qué países son más demandantes de energía eléctrica, en el año más reciente disponible?*
* *¿Qué países son más generados de energía eléctrica, en el año más reciente disponible?*
* *En relación con las preguntas anteriores, ¿cómo cambia la foto si miramos por continente?*
* *¿Qué países se encuentran, en la fecha más reciente, en déficit de energía eléctrica (demandan más de lo que consumen)? ¿Cuáles se hallan en superávit?*
* *¿Cómo se distribuyen los modos de producción de energía, a nivel mundial y continental?*

Ilustración 1. Tipo de gráficos objetivos para las preguntas geográficas



Desde el punto de vista **temporal:**

* *¿Cómo ha evolucionado la demanda, el consumo, y la capacidad de generación de energía eléctrica en los últimos años, a diferentes niveles geográficos (global, continental, nacional)?*
* *¿Cómo han evolucionado estas tres variables en función el modo de generación de dicha energía?*
* *¿Se aprecia un acercamiento cada vez mayor al uso de energías limpias? ¿Depende del país o el continente?*



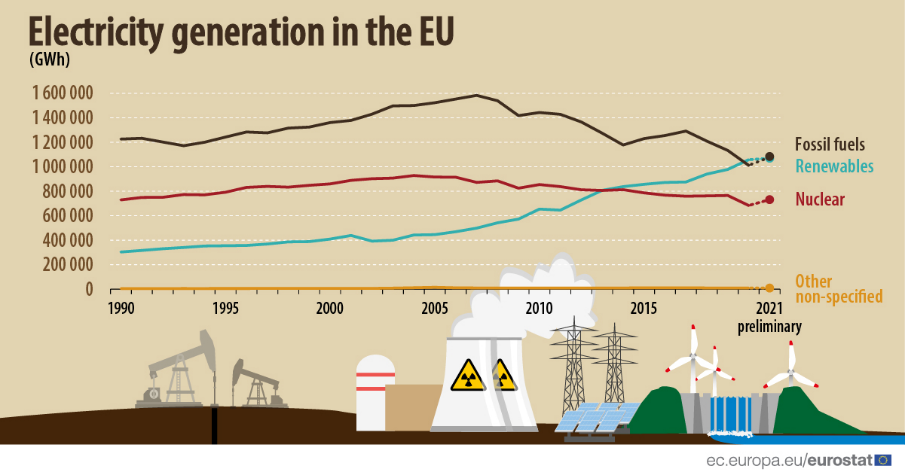


Ilustración 2. Tipo de gráficos objetivo para las preguntas temporales

### Referencias

1. <https://cnnespanol.cnn.com/2022/06/23/alemania-crisis-gas-rusia-trax>
2. <https://wmo.int/es/media/el-cambio-climatico-pone-en-riesgo-la-seguridad-energetica>

1. Para el país usaremos, concretamente, los códigos de tres letra ISO-3166, que se incluyen en ambos conjuntos de datos. [↑](#footnote-ref-1)