Máster Universitario de Ciencia de Datos

Visualización de datos

PRAC2

Alumno: Javier Muñoz Ramón

Enunciado

Parte II: Proyecto de Visualización

En esta segunda parte de la práctica, el estudiante tendrá el reto de desarrollar una visualización de datos que no solo demuestre conocimientos técnicos y teóricos en el campo de la visualización de datos, sino que también integre de manera efectiva las decisiones de diseño necesarias para alcanzar los objetivos planteados. Este proyecto será una oportunidad para consolidar lo aprendido durante la asignatura y poner en práctica habilidades de análisis, diseño y comunicación visual. Este proyecto no solo busca evaluar los conocimientos adquiridos, sino también fomentar la creatividad, la innovación y la capacidad del estudiante para comunicar historias a partir de los datos, utilizando visualizaciones que sean tanto efectivas como impactantes.

Utilizando el conjunto o conjuntos de datos previamente seleccionados y validados en la primera parte de la práctica, se espera que el estudiante implemente las mejoras sugeridas por el profesor, reforzando la calidad del proyecto.

Objetivos principales de la práctica:

1. Elección adecuada de herramientas y técnicas: Identificar y utilizar herramientas específicas para la creación de visualizaciones que se alineen con las características del conjunto de datos, el tipo de análisis realizado y los objetivos del proyecto.
2. Creación de un proyecto profesional: Diseñar una visualización que combine estructura, diseño y contenido de calidad profesional, orientada a públicos especializados y no especializados.
3. Respuestas a preguntas clave: Formular y responder razonadamente. a preguntas clave relacionadas con los datos, utilizando la visualización como un medio para facilitar la exploración y la comprensión.
4. Diseño interactivo y comunicativo: Incorporar elementos interactivos que mejoren la experiencia del usuario y favorezcan la comunicación efectiva de los resultados.

Requisitos adicionales: Además de la visualización, el estudiante deberá presentar un video explicativo que aborde los siguientes aspectos, distribuidos en los porcentajes indicados:

[20%] Proceso de creación: Explicar las etapas del desarrollo, las decisiones de diseño tomadas y los fundamentos detrás de dichas decisiones.

[20%] Presentación en vivo: Mostrar las características de la visualización mientras se navega por ella, destacando aspectos clave del diseño y la funcionalidad.

[15%] Conjunto de datos: Describir brevemente las características más relevantes del conjunto de datos utilizado, su origen y cualquier proceso de preparación realizado.

[20%] Preguntas clave: Detallar las preguntas que la visualización responde y cómo estas se abordan a través del diseño interactivo y analítico.

[15%] Interactividad: Demostrar los elementos interactivos disponibles, explicando cómo contribuyen a la experiencia del usuario. Incluir reflexiones sobre aspectos de accesibilidad.

[10%] Reflexión final: Responder a preguntas como: ¿Qué he aprendido de los datos y de las técnicas empleadas? ¿Qué limitaciones he encontrado? ¿Qué me habría gustado hacer y no pude?

El video deberá tener una duración de entre 4 y 6 minutos. Respetar este rango de tiempo es esencial, ya que se evaluará tanto la capacidad de síntesis como la calidad del guión.

Publicación y entrega del proyecto:

La visualización debe estar disponible públicamente en línea y ser accesible sin necesidad de registro. Puedes usar plataformas como GitHub Pages, Surge.sh, o cualquier otro espacio web accesible. Si se utiliza código, este deberá publicarse en un repositorio público, como GitHub o GitLab, y debe incluirse una licencia de código abierto.

Todos los archivos necesarios para ejecutar la visualización en un servidor web local deben estar incluidos.

Si se emplean herramientas como Tableau, es posible publicarlo en Tableau Public, o bien en plataformas como Observable, Infogram o Flourish, según sea pertinente.

1. Limpieza y selección de variables final

Continuando lo mencionado en el final del trabajo anterior, con la intención de tener un *dataset* mucho más concreto y reducido, analicé de nuevo las correlaciones después de fusionar los tres *datasets* seleccionados.

En la sección donde se analizó la matriz de correlación anterior, se pudieron ver varias combinaciones de variables que, debido a su altísima correlación, parecían indicar colinealidad. Con lo cual, a continuación se mostrarán las modificaciones y los resultados de estas evaluaciones.

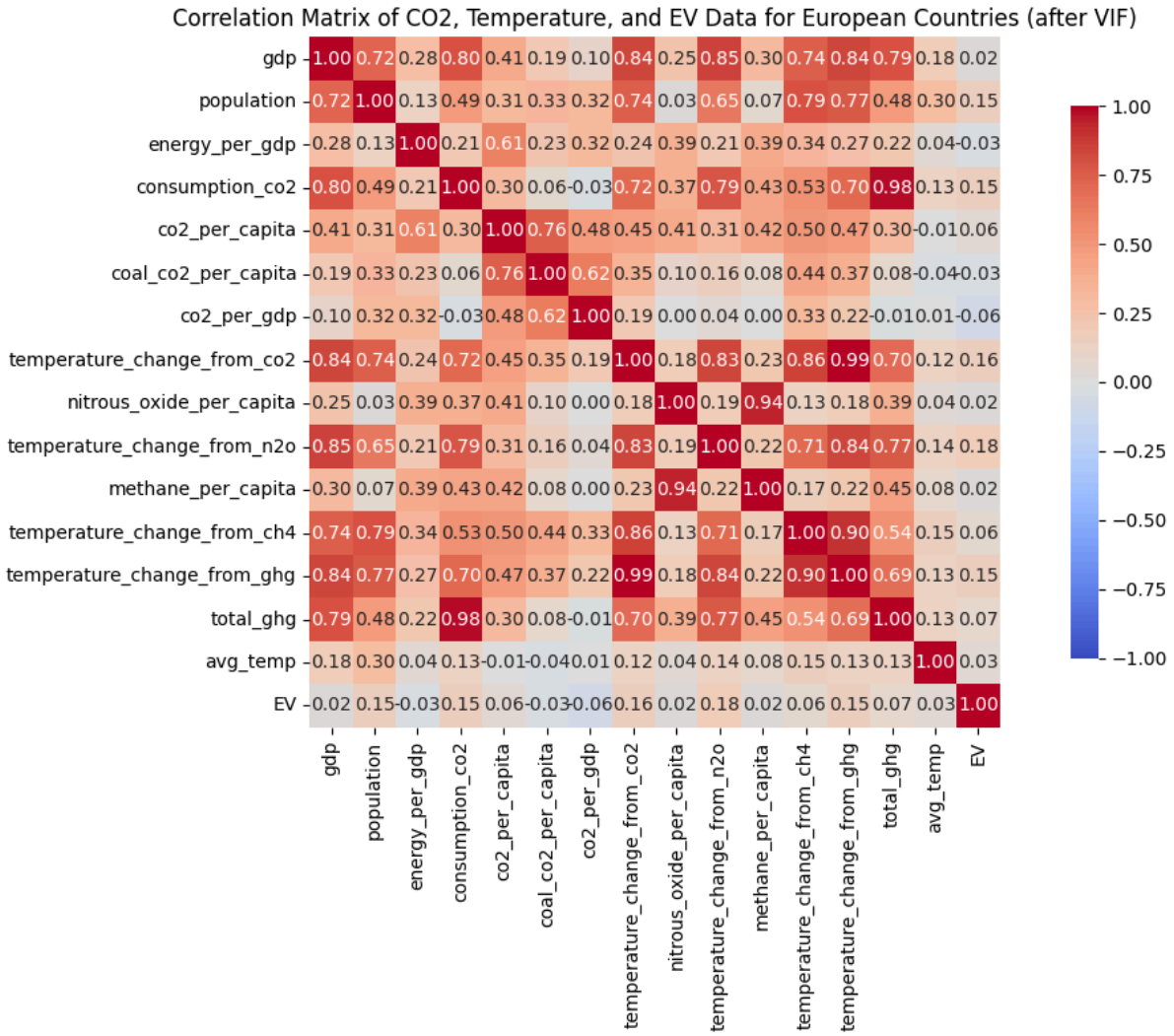
Empezamos haciendo algunas transformaciones, como la de convertir *'year'* a formato *'datetime'*, rellenando los valores vacíos con ceros, y guardando el nuevo *dataframe* en una variable que solo contemple las variables numéricas para calcular la correlación:

```
# We convert year to datetime64:
data_co2_temps_ev_eu_merged['year'] = pd.to_datetime(data_co2_temps_ev_eu_merged['year'], format='%Y')

# We convert all NaN values to 0:
data_co2_temps_ev_eu_merged.fillna(0, inplace=True)

# We calculate the correlation matrix of the merged DataFrame with the numerical data using the numerical = True parameter:
data_co2_temps_ev_eu_corr = data_co2_temps_ev_eu_merged.corr(numeric_only=True)
```

Es importante remarcar que la decisión de rellenar con ceros es más porque en aquellos valores faltantes solamente se trataba de aquellos que no tenían ningún registro para ese país y año concreto, ya que no todos los países que escogimos tenían, por ejemplo, registros de consumo de CO_2, en especial en el principio de los datos que databan en los comienzos del siglo XVIII, además de los años donde todavía no existían o se comercializaban los vehículos eléctricos.

Al ser la mayoría de cálculos que se realizarán medias de las sumatorias, estos ceros no deberían afectar a los resultados, siempre que existan otros registros que podamos calcular.

Con Matplotlib y Seaborn graficamos la matriz de correlación:



Vemos las variables con muy alta correlación y probable colinealidad en esta matriz, pero para tener un cálculo más informativo sobre la colinealidad, utilizaremos el VIF con la librería Statsmodel:

```python
# We select the numerical columns for the VIF calculation:
data_co2_temps_ev_eu_merged_num = data_co2_temps_ev_eu_merged.select_dtypes(include='number')

# We add a constant to the DataFrame for the VIF calculation:
data_co2_temps_ev_eu_merged_num = sm.add_constant(data_co2_temps_ev_eu_merged_num)

# We calculate the VIF for each variable:
vif_data = pd.DataFrame()
vif_data['Variable'] = data_co2_temps_ev_eu_merged_num.columns
vif_data['VIF'] = [variance_inflation_factor(data_co2_temps_ev_eu_merged_num.values, i) for i in range(data_co2_temps_ev_eu_merged_num.shape[1])]

# We show the VIF data:
vif_data = vif_data[vif_data['Variable'] != 'const'].sort_values(by='VIF', ascending=False).reset_index(drop=True)
vif_data.round(2)
```

27] ✓ 0.0s

	Variable	VIF
0	temperature_change_from_ghg	234.96
1	temperature_change_from_co2	160.79
2	consumption_co2	28.81
3	total_ghg	27.10
4	methane_per_capita	10.74
5	nitrous_oxide_per_capita	9.76
6	gdp	7.98
7	temperature_change_from_ch4	7.08
8	co2_per_capita	6.54
9	temperature_change_from_n2o	5.64
10	coal_co2_per_capita	5.04
11	population	4.45
12	energy_per_gdp	2.41
13	co2_per_gdp	1.97
14	EV	1.37
15	avg_temp	1.18

Podemos observar que obtenemos variables con un VIF superior a 10 en:

- ***'temperature_change_from_gh'*** -> 234,96 -> Casi totalmente explicada por otras variables
- ***'temperature_change_from_co2'*** -> 160,78 -> Altamente redundante con otras variables
- ***'consumption_co2'*** -> 28,80 -> Muy fuerte colinealidad
- ***'total_ghg'*** -> 27,09 -> Muy fuerte colinealidad
- ***'methane_per_capita'*** -> 10,74 -> Límite aceptable pero alto

De todas las variables que aparecen en esa lista, y a fin de evitar la colinealidad y el exceso de variables superfluas, decidí eliminar estas:

- ***'consumption_co2'***
- ***'temperature_change_from_co2'***
- ***'temperature_change_from_ch4'***
- ***'temperature_change_from_n2o'***
- ***'coal_co2_per_capita'***
- ***'methane_per_capita'***
- ***'nitrous_oxide_per_capita'***

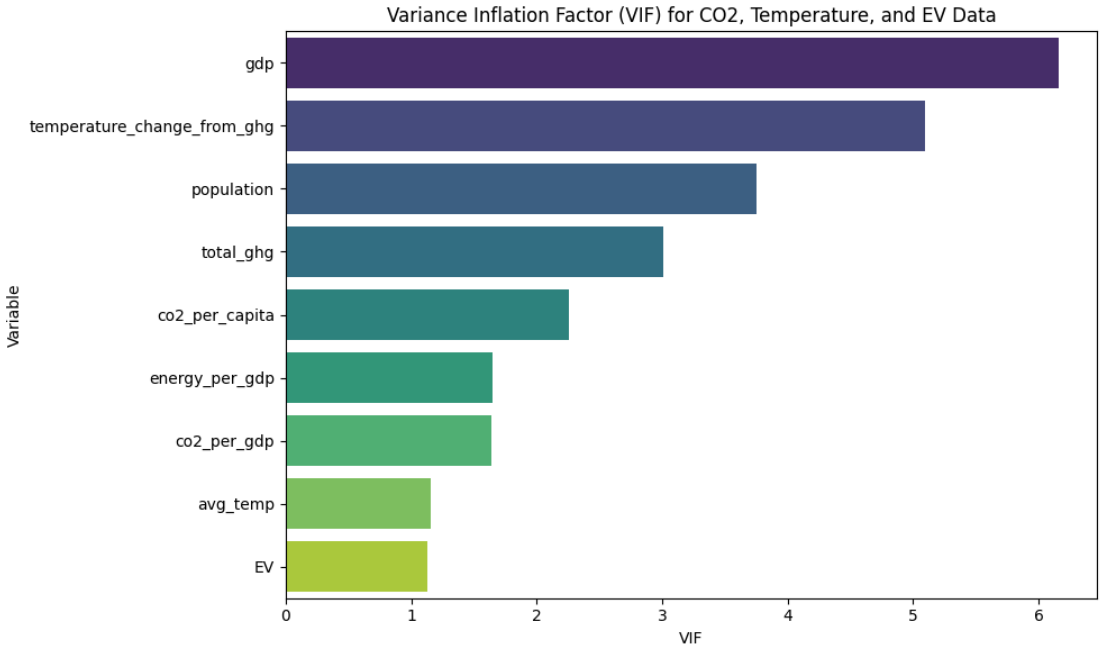
Teniendo en cuenta que *'total_ghg'* ya debería explicar *'consumption_co2'* al incluir todos los gases de efecto invernadero (ghg = *greenhouse gases*), y que *'temperature_change_from_ghg'* hace lo mismo con los otros *'temperature_change_from…'*, pensé que estas variables no serían esenciales y que eliminarlas podría reducir mucho la colinealidad.

También decidí eliminar *'coal_co2_per_capita'* porque no me interesaba y creía que podía haber un exceso de variables. Y finalmente *'methane_per_capita'* y *'nitrous_oxide_per_capita'*, ya que tenían un VIF alto, y me quedé con *'co2_per_capita'* al tenerlo inferior y ser el gas de efecto invernadero más abundante y el que más se estudia respecto al calentamiento global.

Después de eliminar esas variables, volvemos a calcular el VIF y obtenemos valores mucho mejores:

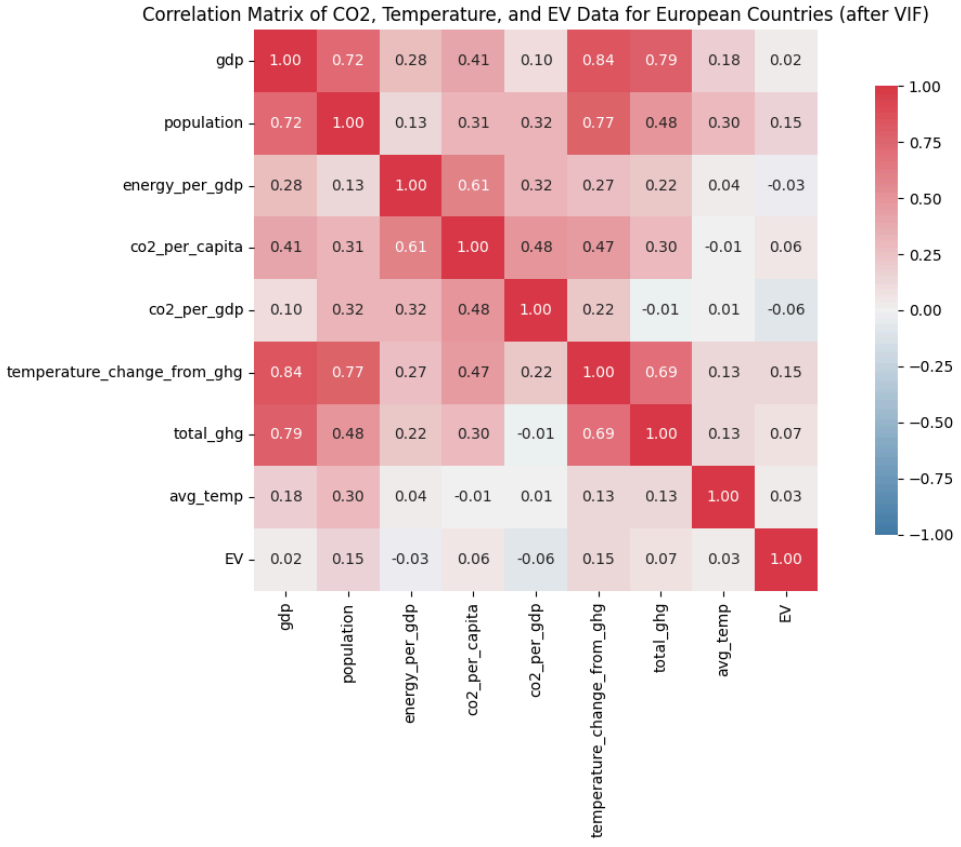
	Variable	VIF
0	gdp	6.16
1	temperature_change_from_ghg	5.10
2	population	3.75
3	total_ghg	3.01
4	co2_per_capita	2.26
5	energy_per_gdp	1.65
6	co2_per_gdp	1.64
7	avg_temp	1.15
8	EV	1.13

Lo graficamos:



Y todos están por debajo de 5, que es mucho mejor que estar alrededor de 10 o por encima.

Hacemos la matriz de correlación después de esta eliminación:



Así vemos los datos sin colinealidad y mucho más simplificados.

Como hemos estado operando con una matriz totalmente numérica para trabajar con las correlaciones y el VIF, añadimos a este *dataframe* las variables categóricas pertinentes y algunas transformaciones de tipo de dato, limpieza y demás para poder crear los gráficos en la siguiente sección:

```
# We add the 'year' and 'country' columns to the numerical DataFrame and call it `data_co2_temps_ev_eu_final`:
data_co2_temps_ev_eu_final = data_co2_temps_ev_eu_merged_num.copy()
data_co2_temps_ev_eu_final['year'] = data_co2_temps_ev_eu_merged['year']
data_co2_temps_ev_eu_final['country'] = data_co2_temps_ev_eu_merged['country']

# We also delete the 'const' column:
data_co2_temps_ev_eu_final = data_co2_temps_ev_eu_final.drop(columns='const')

# We change the 'avg_temp' to 2 decimal places:
data_co2_temps_ev_eu_final['avg_temp'] = data_co2_temps_ev_eu_final['avg_temp'].round(2)

# We convert 'EV' to int64:
data_co2_temps_ev_eu_final['EV'] = data_co2_temps_ev_eu_final['EV'].astype('int64')

# We make sure that the 'year' only shows the year, not the month and day:
data_co2_temps_ev_eu_final['year'] = data_co2_temps_ev_eu_final['year'].dt.year

# We drop the gdp column because it has a high correlation with 'population' and 'energy_per_gdp':
data_co2_temps_ev_eu_final = data_co2_temps_ev_eu_final.drop(columns='gdp')
```

2. Proceso de creación de las visualizaciones

En lo referente al proceso de creación de las visualizaciones, decidí tomar la idea del trabajo anterior que trataba de, primeramente, pensar en qué información geográfica podría representar como mapa coroplético.

A medida que exploraba los datos del *dataframe*, iba pensando qué elementos podía incluir, entre los que creí que quizás lo más representativo de lo que añadí al enriquecer los datos eran las temperaturas. Con estas, pude representar las medias en función de rangos de años de una forma más visual y fácil de entender.

Para el resto de visualizaciones, el tipo de gráfico que escogí fue el de líneas o *line graph*. Con estos, me parecía mucho más sencillo representar las tendencias de diferentes variables con una serie temporal en un eje cartesiano. De este modo, podemos ver la influencia y la correlación (aunque no sea exactamente la causalidad), de los aumentos y los descensos de sus valores. Para asegurar que estos se representaban de forma que podían compararse entre sí, creí necesario normalizar todas las variables numéricas.

Como método de creación de las visualizaciones decidí utilizar Python como Matplotlib, Seaborn y Folium para graficar, y Plotly para crear elementos interactivos en ellas.

3. Preguntas clave para el desarrollo del proyecto

En la primera parte del trabajo anterior nos propusimos estas preguntas:

- ¿Los países con mayor cuota de mercado de vehículos eléctricos reducen su CO_2 per cápita más rápido que el resto?
- ¿Qué países tienen una alta adopción de vehículos eléctricos, pero no han visto reducidas sus emisiones de CO_2 de manera significativa?
- ¿Los países que reducen el CO_2 se calientan menos?
- ¿Existe un punto de inflexión donde la adopción de vehículos eléctricos marca una reducción importante de los gases de efecto invernadero?
- ¿Qué tan rápido aumentan las temperaturas a la par que aumentan las emisiones?

Para continuar con esto, analicemos brevemente cuáles son las mayores fuentes de cada uno de los tres gases de efecto invernadero que existían en el *dataset* original, y por qué hemos elegido el CO_2 como el más importante:

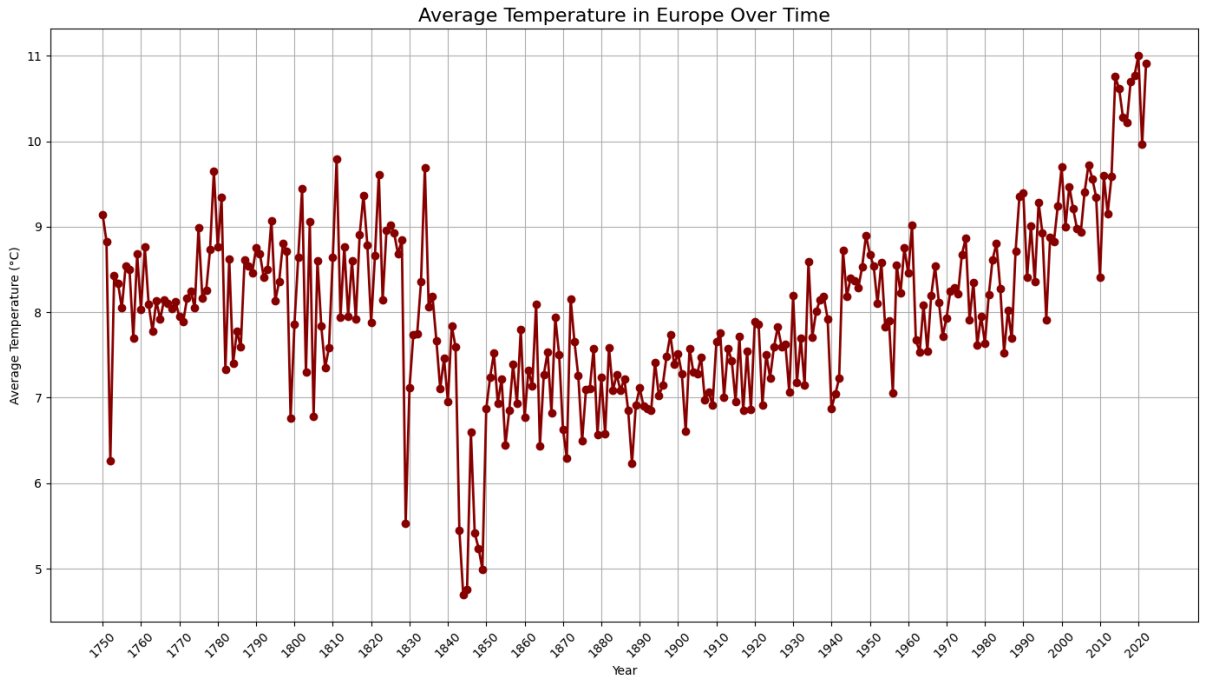
- **CO_2**: lo producen mayormente los combustibles fósiles, el carbón, el petróleo y el gas natural. Es el responsable de más o menos el 75% del calentamiento global y permanece siglos en la atmósfera. Es el más abundante.
- **Metano (CH_4)**: suele producirlo mucho más la agricultura y el ganado, aunque también al extraer petróleo y gas, y la descomposición de residuos. Es 25 veces más potente que el CO_2 y dura solo de 10 a 12 años. Es menos abundante que el CO_2.
- **Óxido nitroso (N_2O)**: lo producen los fertilizantes agrícolas y el procesamiento de las aguas residuales. Es mucho más potente que el CO_2 en referencia al calentamiento (298 veces más). Dura alrededor de 114 años en la atmósfera y es el menos abundante.

Dado que el CO_2 es el más abundante y es el que producen los vehículos en mayor medida, escogí las variables *'total_ghg'* y *'co2_per_capita'* para trabajar en la respuesta de las dudas que se formularán, junto con las de *'avg_temp'* y *'EV'*.

Ya que el *dataset* principal que hemos utilizado para realizar el análisis y las visualizaciones es el centrado en la contaminación con los gases de efecto invernadero, y los que hemos utilizado para enriquecer ese mismo son los de temperaturas por país y los vehículos eléctricos, la pregunta final debería ser: ¿qué influencia tiene el nuevo auge de los vehículos eléctricos en la reducción de la contaminación, y la consecuente reducción de las temperaturas medias?

Esta pregunta por sí misma requiere una exploración de los datos que hace formular varias preguntas antes de llegar a la conclusión de la misma. Primeramente, ¿hay un aumento de las temperaturas medias de Europa?

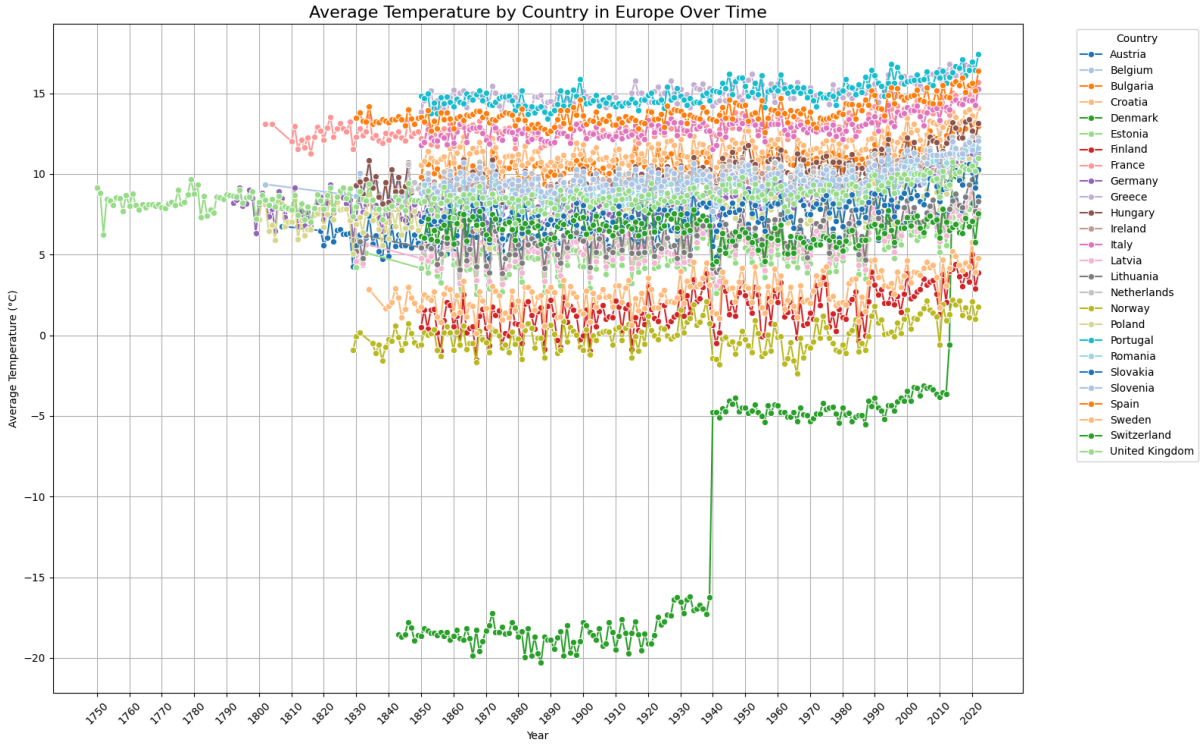
Esto se explica de forma sencilla con el primer gráfico:



Evidentemente, excepto por alguna fluctuación concreta, las temperaturas medias tienen una tendencia al aumento desde mediados del siglo XIX, que es cuando comenzó la revolución industrial.

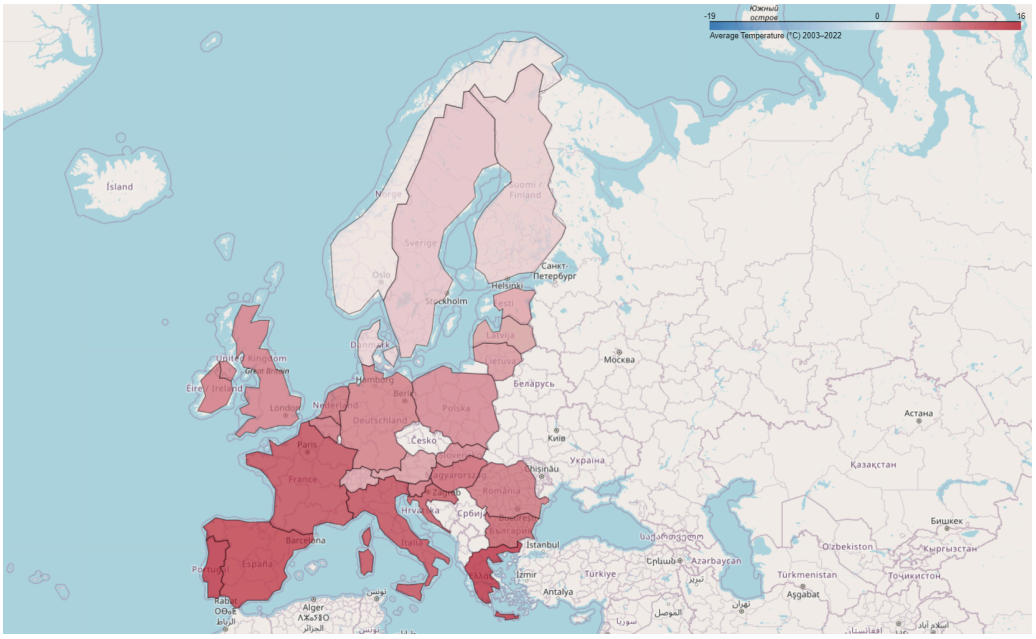
Otra pregunta sería: ¿se puede observar la misma tendencia en todos los países?

Esto se comprueba fácilmente con otro gráfico de líneas que incluya todos los países que seleccionamos desde el principio:



Excepto el caso de Dinamarca, que muy probablemente sean *outliers* de valores erróneos mal registrados, vemos que el resto de países siguen una tendencia exactamente igual que la media de Europa, siendo algunos más, y otros menos calurosos por motivos geográficos.

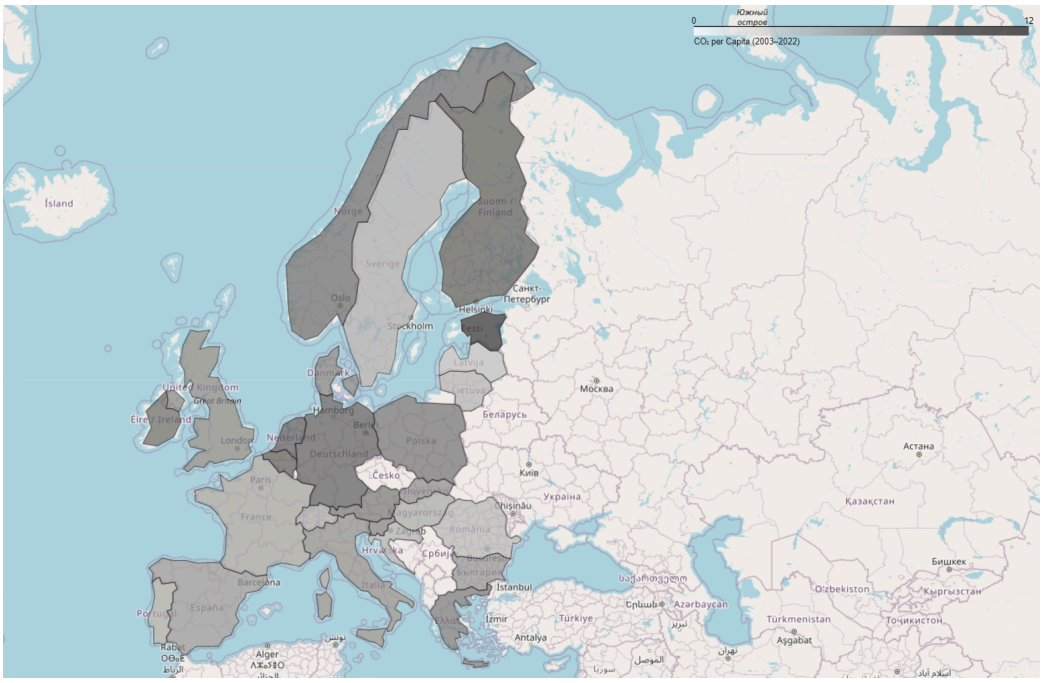
Respecto a las temperaturas medias, se han creado también dos mapas coropléticos comparativos del primer siglo y medio del *dataset*, y luego otro de los últimos 20 años para hacer comparativas. Ejemplo del mapa (se explica mejor en el vídeo):



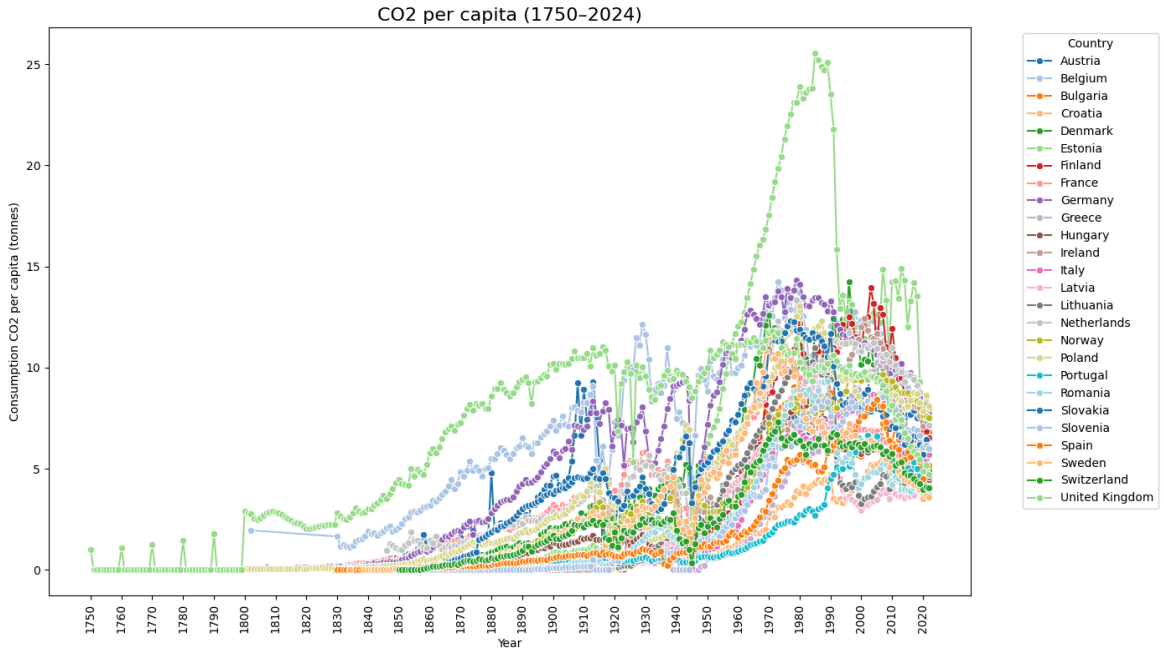
Para responder esto debemos tener en cuenta factores importantes como la posición geográfica.

La pregunta que nos podemos hacer con los siguientes datos agregados es: ¿todos los países contaminan igual per cápita?

Esta pregunta la podemos ver claramente con otra versión del mismo mapa coroplético, pero dedicado al CO_2 per cápita, para verlo más claramente (se explica en el vídeo):



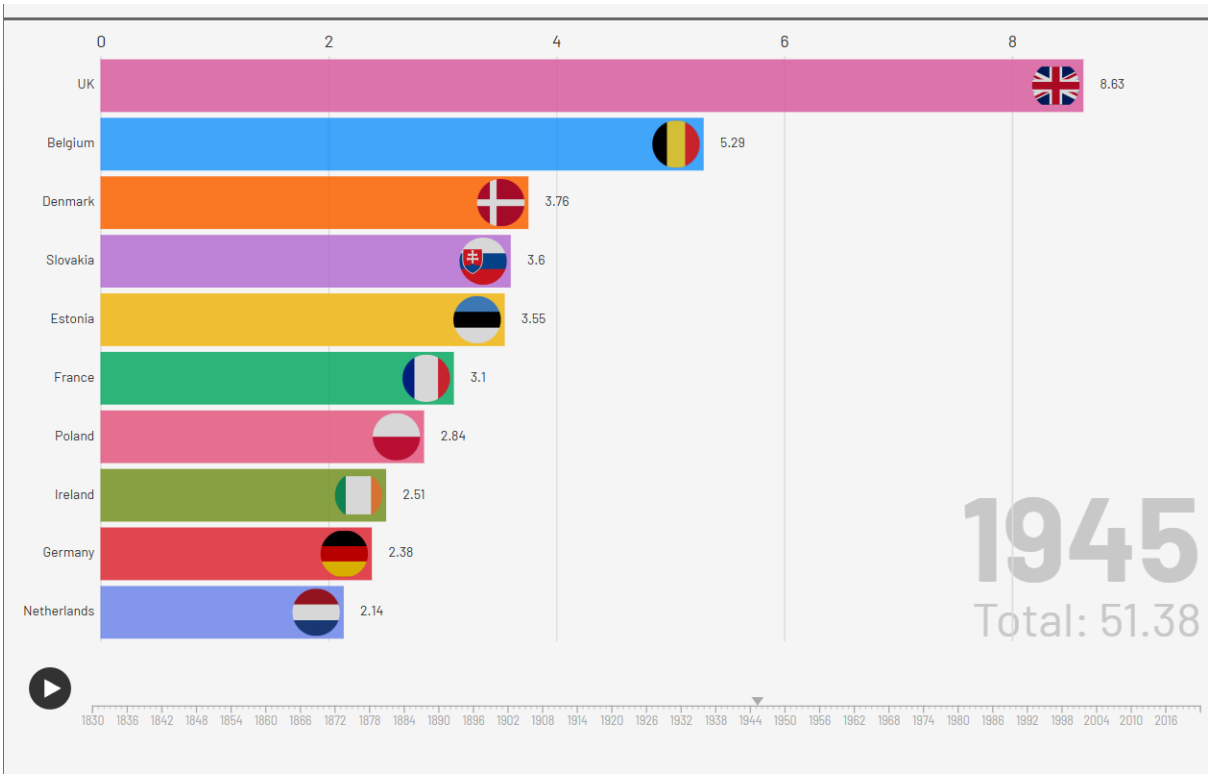
También podemos visualizar con otro gráfico de líneas, las tendencias de contaminación de CO_2 per cápita desglosadas por países para ver si existen grandes aumentos y disminuciones en algunos países:



Todos siguen una tendencia muy similar, pero Gran Bretaña tuvo un aumento mucho mayor durante una parte importante del siglo XX. Eso fue debido a la necesidad de reconstrucción del país posteriormente a la Segunda Guerra Mundial, y por el exceso de uso del carbón como combustible, además del crecimiento industrial, lo cual causó eventos como la Gran Niebla de Londres de 1952, donde murieron miles de personas por la toxicidad del aire por los productos de la combustión.

Para explicar estos datos de una forma más dinámica, decidí utilizar la famosa *bar race chart* con Flourish. Y para ello tuve que pivotar los datos del co2 per cápita, haciendo que el dataset esté compuesto por una columna de países, y el resto con columnas de año que contienen los valores del CO2 per cápita, y otra añadida con enlaces a las imágenes de las banderas.

Este fue el resultado:

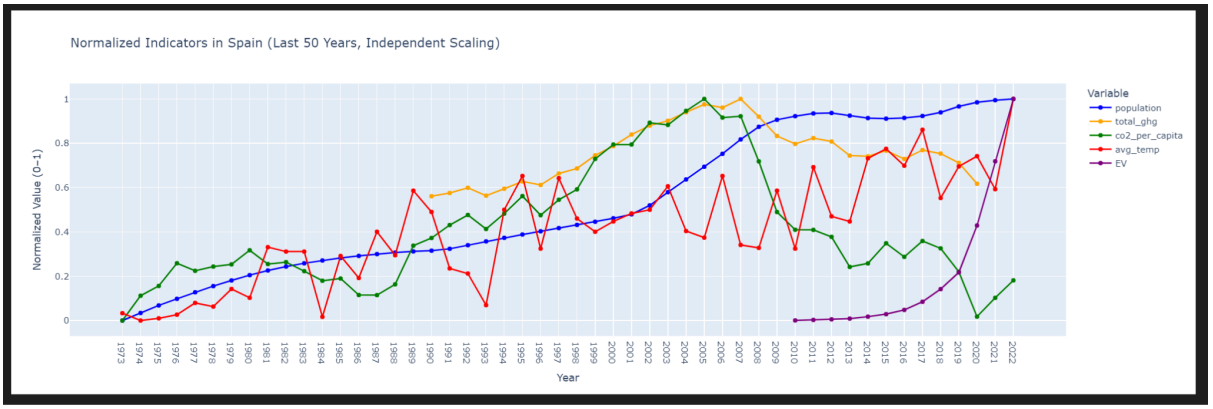


Este tipo de gráficos son muy llamativos porque representan los datos de forma animada y nos permiten visualizar las tendencias de un modo más ameno que si analizamos la información en un gráfico como el anterior, donde la cantidad de datos que hay se solapan unos con otros y es complicado saber qué países estamos viendo entre toda la maraña de líneas que se forman.

Por último, me planteé la duda final, con algunos matices. Añadiendo las variables de población, de gases de efecto invernadero totales, de temperatura media y de vehículos eléctricos, comencé con las dudas:

- ¿Está la población correlacionada con el aumento del CO_2 per cápita?
- ¿Siguen el CO_2 per cápita y los gases de efecto invernadero totales las mismas tendencias?
- ¿La tendencia de las temperaturas medias se ve afectada por el aumento y la reducción de los gases de efecto invernadero y el CO_2 per cápita?
- ¿Cómo afecta la aparición y el auge de los vehículos eléctricos a todas estas variables?
- Y, por último, ¿se puede extrapolar el resultado de un país como España a toda Europa?

Todo esto se explica con más detalle en el vídeo con dos gráficos como el de este ejemplo:



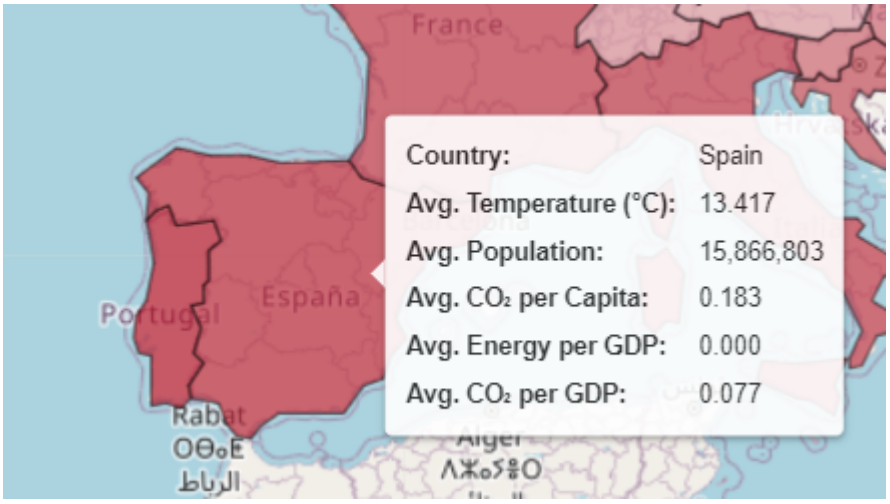
Estas visualizaciones completas se pueden ver en este enlace de GitHub Pages:

https://jmura84.github.io/VISUALIZACION_DATOS/

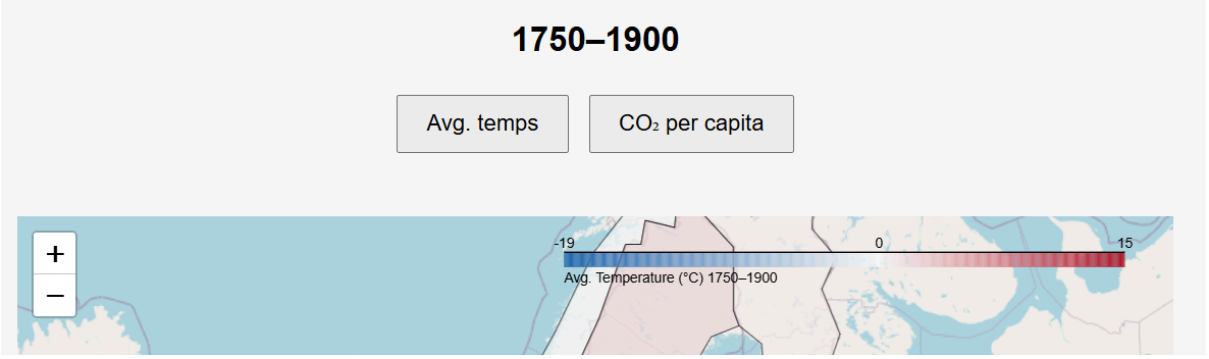
4. Interactividad de las visualizaciones

En gráficos como los primeros, donde vemos las tendencias de temperaturas, o el de tendencias de CO_2 per cápita y por país decidí que no era necesario añadir ninguna interacción, ya que quería que, a pesar de que algunos pareciesen “caóticos” a primera vista, como eran bastante explicativos por sí mismos, estos ya aportaban la información que deseaba mostrar.

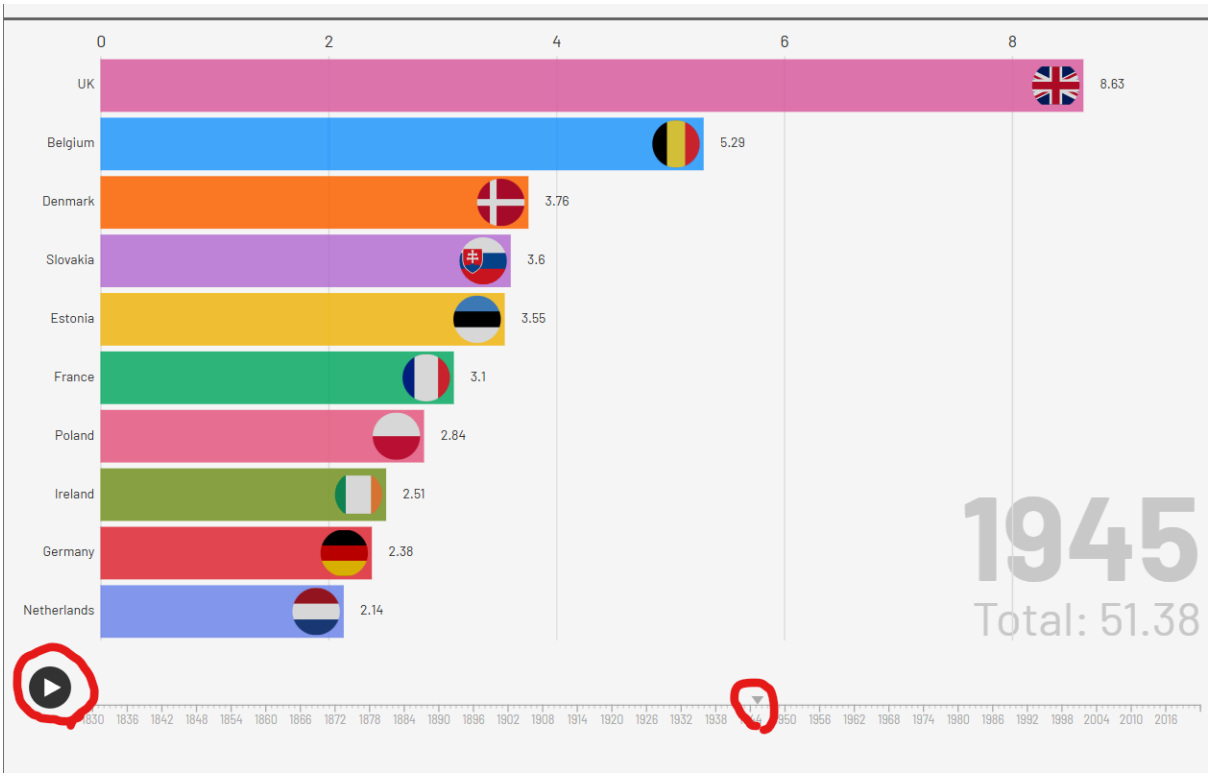
En casos como el de los mapas coropléticos sí que decidí que sería interesante poder interactuar con cada país haciendo *hover* sobre cada uno, y que apareciese un *pop-up* donde se desglosaba la información de ese país en ese rango de años, para que podamos ver realmente qué factores podían intervenir, o qué datos podíamos analizar. Por ejemplo:



Y en el HTML donde he enlazado todas las visualizaciones, también he añadido botones para cambiar del mapa coroplético de temperaturas al del CO_2:



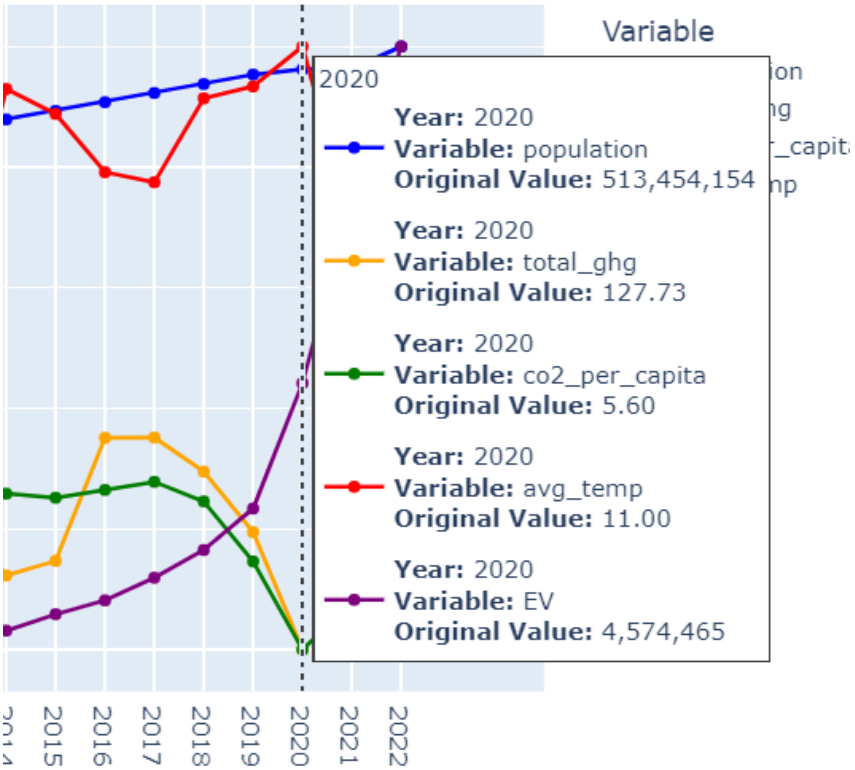
En cuanto al gráfico de Flourish, lo único interactivo que hay es el botón de play para que comience la animación, y la posibilidad de mover el puntero que hay en el eje de los años como si fuese un *slider*:



El único defecto que le veo es que la animación comienza automáticamente cuando hago *scroll* por la web en HTML que hemos creado. Hubiese sido mejor poder presionar el botón en el momento en que yo desee que comience la animación.

Respecto a los gráficos de líneas que comparan las variables que responden las preguntas finales, decidí que ahí sí sería interesante también disponer de un *pop-up* al hacer *hover* sobre cada año, para así obtener el valor que contiene cada uno. La única diferencia entre ambos gráficos es que el de España contiene los valores por año tal cual, y el de Europa tiene los datos modificados, como, por ejemplo, la población y los vehículos eléctricos son una suma de todos los países, y la de total de gases de efecto invernadero, las temperaturas y el CO_2 per cápita son las medias.

Ejemplo:



5. Conclusiones

Para concluir, solo decir que este trabajo de esta asignatura me ha sido muy útil para comprender cómo partiendo de unos datos, podemos ser capaces de ir pensando dudas que nos van surgiendo, para luego encontrar preguntas que nos conduzcan a nuevos datos que, una vez agregados, nos serán útiles para encontrar información que nos puede ser de gran utilidad.

Además, no solo trata de encontrar ese valor intrínseco en cada pregunta respondida, sino en ser capaces de plasmar aquello que queremos conocer con unas visualizaciones que sean adecuadas para entender los datos de un modo lo más sencillo posible, y con el máximo de información condensada en poco más de un golpe de vista y, en ocasiones, un poco de exploración.

Finalmente, lo que también hay que tener muy en cuenta es conseguir transmitir esas respuestas a un público de modo que tenga sentido, coherencia, y que pueda ser interesante, siendo también muy importante hilar la narración cohesionadamente, y que ayude a los demás a dar una perspectiva diferente de la que puedan sacar sus propias conclusiones.