Máster Universitario de Ciencia de Datos

Visualización de datos

PRAC1

Alumno: Javier Muñoz Ramón

Enunciado

Parte I: selección del conjunto de datos

Esta actividad, primera parte de la práctica final, consiste en la selección por parte del estudiante de un conjunto de datos de su interés que será usado en el proyecto de creación de la visualización de datos, de acuerdo con unos criterios establecidos. Básicamente, la temática es libre, pero se valorarán los aspectos siguientes:

[10 %] Justificad brevemente vuestra selección, sea por motivos personales o profesionales.

[10 %] La relevancia del conjunto de datos en su contexto. ¿Son datos actuales? ¿Tratan un tema importante por algún colectivo concreto? ¿Se ha tenido en cuenta la perspectiva de género?

[25 %] La complejidad (medida, variables disponibles, tipos de datos, etc.). Debe tener del orden de miles de registros mínimo y debe tener un mínimo del orden de decenas de variables. ¿Combina datos categóricos y cuantitativos? ¿Incluye otros tipos de datos? La riqueza en tipología de variables os puede ayudar a realizar un trabajo más brillante: valores discretos, continuos, fecha u hora, lógicos, cartográficos.

[25 %] La originalidad. Se valora no repetir los conjuntos de datos clásicos o muy trabajados (https://medium.com/@dhruval_/50-popular-datasets-for-your-next-data-science-project-11550daf8d2) ni temas ya muy tratados (p. ej. Covid-19, tráfico, criminalidad…). Podéis combinar o mejorar el conjunto de datos. En el primer caso, enriquecer el conjunto de datos con otros diferentes para dar un enfoque nuevo. En el segundo caso, generando nuevas métricas e indicadores con las variables existentes mediante transformaciones. ¿Hay otras visualizaciones basadas en este conjunto de datos? ¿Es una evolución o actualización de un conjunto anterior? ¿Habéis enriquecido un conjunto de datos ya existente?

[30 %] Las cuestiones que responderéis con la visualización de datos, ¿tienen en cuenta los puntos anteriores? ¿Han sido planteadas en otras visualizaciones u otros proyectos? ¿Son adecuadas para el conjunto de datos elegido? En este punto, elaborar un diccionario de las variables, su significado y si es un hecho a estudiar o una dimensión que lo mide, os puede ayudar.

1. Justificación de la selección

El tema elegido para esta práctica es el de la relación entre las emisiones de CO_2 y gases de efecto invernadero con los aumentos de temperatura regionales. Específicamente en varios países europeos

Los motivos de la selección son, más que profesionales, debido a que mi intención es que uno de los gráficos sea de tipo coroplético, para poder visualizar los datos de una forma más visual y atractiva, y luego crear gráficos diversos que requieran una representación de datos numéricos y categóricos de un modo más clásico, por ejemplo, en un eje cartesiano, o incluso menos convencional. Además, es un tema que da bastante pie a poder agregar datos de otros *datasets* para, de este modo, buscar correlaciones que podrían ser interesantes.

Como la polución y el calentamiento global son temas que tienen muchos estudios, disponemos de muchas variables con las que podemos jugar y ver si nos sirven para extraer información de valor que nos permita, quizás, poder deducir o inferir alguna conclusión más visualizando los datos más relevantes del *dataset* original y los agregados.

Los *datasets* escogidos de momento son:

- **Global CO_2 and Greenhouse Gas Emissions:** https://www.kaggle.com/datasets/mexwell/global-co2-and-greenhouse-gas-emissions
- **Earth Temperature Data By Country (1743 - 2024):** https://www.kaggle.com/datasets/anastasiaalyoshkina/earth-landsurface-temperature-data-1750-2024
- **Global Electric Vehicle Sales Data (2010-2024):** https://www.kaggle.com/datasets/muhammadehsan000/global-electric-vehicle-sales-data-2010-2024/data

El primero es en el que se fundamentará la mayor parte del análisis principal de datos, y con los otros se intentará realizar agregaciones de variables y enriquecimiento que podrá ser útil en el momento de extraer ideas que puedan ser interesantes a la hora de buscar interacciones y correlaciones.

2. Relevancia del conjunto de datos

En lo referente a la relevancia del conjunto de datos principal, que es **Global CO_2 and Greenhouse Gas Emissions**, podemos observar que su rango de fechas va desde 1750 a 2022. Se ha descargado desde la página Kaggle, original de Our World in Data (documentación sobre la creación del *dataset* en GitHub: https://github.com/owid/co2-data), y su licencia es Attribution 4.0 International de Creative Commons.

Respecto al segundo (**Earth Temperature Data By Country**), el rango de fechas que contiene es del año 1743 al 2024. También se ha descargado de Kaggle y su licencia es CC0 1.0 Universal de Creative Commons.

El último (**Global Electric Vehicle Sales Data**), tiene un rango de años más corto, que es del 2010 al 2024, y se extiende al 2035, dado que se ha intentado hacer una predicción de la adopción de vehículos eléctricos a futuro. También se ha descargado de Kaggle, y su licencia es Attribution 4.0 International de Creative Commons.

Podríamos decir que, mayormente, el rango de años de cada *dataset* está bastante bien actualizado, y que los rangos de los *datasets* del CO_2 y del de temperaturas cubren la transición que se hizo durante la revolución industrial, en la que se aumentó drásticamente el uso de combustibles fósiles como el carbón, la gasolina, el gas, etc, hasta la actualidad.

El *dataset* de los vehículos eléctricos, sin embargo, solamente cubre 14 años, ya que es la fecha en la que se comenzaron a adoptar, de una forma más masificada, vehículos de muchos tipos con motores que usaban esa tecnología.

Todos ellos se pueden enlazar entre sí por año y por país.

3. Complejidad de los datos

Si analizamos el *dataset* principal, este dispone de 79 variables, de entre las cuales podemos obtener datos categóricos como el país, el año (por no ser un valor con el que hacemos cálculos) y el código ISO del país. Y entre las variables numéricas y continuas tenemos multitud de diferentes datos sobre cada uno de los gases de efecto invernadero, entre los que se encuentran el CO_2, el óxido nitroso y el metano, además de cálculos más generales de los efectos generales de los gases de efecto invernadero combinados.

También podemos ver datos más relacionados con cada país, como son su PIB y la población, que nos pueden ayudar a averiguar si existe correlación entre esos dos datos y la producción de esos gases.

En el *dataset* sobre las temperaturas disponemos de las variables categóricas del país analizado, y el año y el mes cuando se registraron esos datos. Y en las variables numéricas tenemos la temperatura ajustada que se ha registrado en ese país y esa fecha.

Y, finalmente, en el *dataset* sobre la adopción de vehículos eléctricos, vemos las variables categóricas de región geográfica, categoría del dato, el parámetro o característica de cada dato, el método de colección del dato, el tipo de motor de propulsión que usa el vehículo, el año, y la unidad en la que se cuenta en el último valor, que es el numérico.

Debido a la complejidad, por el número de variables del *dataset* original, y por la cantidad de variables categóricas del último y de los distintos tipos de valores de la variable *'value'*, se tendrán que reducir las dimensiones, hacer transformaciones, e incluso elegir variables cuidadosamente para poder hacer un análisis interesante al final, que nos permita hacer las visualizaciones que se requieren.

4. Originalidad de los datos seleccionados

Respecto a la originalidad de los datos seleccionados y de su tema, hay que reconocer que quizás es un tema del que se ha hablado mucho en diversos estudios, pero, como quería que al menos una de las visualizaciones fuese un mapa coroplético, me pareció buena idea escoger el calentamiento global con la polución de los distintos gases de efecto invernadero, seleccionados por año y por país, y luego poder enriquecer los datos con los de la adopción de vehículos eléctricos, ya que aunque no sea la única variable para reducir las emisiones, sí que puede ayudar a comprender un poco mejor el compromiso de los países seleccionados en el estudio con el medio ambiente, y para saber si esa adopción se corresponde con la reducción de emisiones.

Sí, existen visualizaciones sobre la polución emitida por cada país y la correlación con el calentamiento del planeta, y algunos sobre el impacto de los vehículos eléctricos en este tema, pero me pareció interesante poder hacer yo mismo las combinaciones de variables, crear nuevas, o incluso transformar algunas si es necesario. De este modo, quizás se pueden obtener unas nuevas perspectivas al respecto, y a la vez crear unas visualizaciones que no se hayan hecho previamente, con combinaciones de variables que no sean iguales que las de los estudios existentes.

En lo referente a si se ha enriquecido un conjunto ya existente, la idea principal es esa. En el *dataset* principal, la mayor parte de los datos son sobre emisiones, desglosados de muy diversas formas, y luego tenemos también datos más de tipo económico o poblacional, y algún dato en la variación de temperatura basándonos en cada gas. El *dataset* de las temperaturas permite enriquecer el primero proporcionando las temperaturas por mes y por año, del cual podemos sacar una media anual y quedarnos con ese dato al agregarlo al principal. Y del de los vehículos eléctricos podemos obtener, transformando múltiples datos, los vehículos de ese tipo que se han vendido por cada país.

Para simplificar, del primer *dataset* se han filtrado los datos por una lista de países europeos, y también por ciertas variables concretas de cada tipo de gas. De este modo podemos reducir la dimensionalidad y buscar mejor las correlaciones, además de mejorar las visualizaciones en la siguiente parte del trabajo.

5. Cuestiones a responder y desglose de variables

Las cuestiones que podríamos responder, serían:

- ¿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?
- Etc.

Como se han reducido el número de países y variables para facilitar el análisis y las transformaciones previas, a continuación mostraré la selección de datos que he escogido de cada *dataset* hasta el momento, con posibilidad de cambios en la segunda parte del trabajo en función de si algunos datos pueden ser más o menos interesantes o útiles.

Global CO_2 and Greenhouse Gas Emissions

Lista de países escogidos (solo de Europa):

'Spain', 'France', 'Germany', 'Italy', 'Netherlands', 'Belgium', 'Austria', 'Sweden', 'Norway', 'Finland', 'Denmark', 'United Kingdom', 'Greece', 'Portugal', 'Ireland', 'Switzerland', 'Poland', 'Czech Republic', 'Hungary', 'Slovakia', 'Romania', 'Bulgaria', 'Croatia', 'Slovenia', 'Lithuania', 'Latvia', 'Estonia'.

Variables categóricas escogidas:

- country: país donde se mide la temperatura
- year: año en el que se mide la temperatura

Variables numéricas escogidas:

- gdp: PIB del país analizado
- population: población del país
- energy_per_gdp: medida del PIB por consumo de energía primario
- consumption_co2: consumo de CO_2
- co2_per_capita: CO_2 calculado per cápita
- coal_co2_per_capita: CO_2 generado por carbón per cápita
- co2_per_gdp: CO_2 calculado con base en el PIB
- temperature_change_from_co2: cambio de temperatura calculado con el CO_2
- nitrous_oxide_per_capita: óxido nítrico generado per cápita
- temperature_change_from_n2o: cambio de temperatura calculado con el óxido nítrico
- methane_per_capita: metano generado per cápita
- temperature_change_from_ch4: cambio de temperatura generado por el metano
- temperature_change_from_ghg: cambio de temperatura causado por los gases de efecto invernadero en general
- total_ghg: gases de efecto invernadero totales

Earth Temperature Data By Country (1743 - 2024)

Variables categóricas:

- Year: año de registro de los datos de temperatura
- Month: mes de registro
- Country: país de registro

Variable numérica:

- Adjusted Temperature: cálculo de temperatura ajustada para corregir las diferencias en los métodos de medición y la influencia moduladora del océano o el mar

Global Electric Vehicle Sales Data (2010-2024)

Variables categóricas:

- region: país de la observación
- year: año del registro
- unit: unidades que se cuentan en ‘value’ (porcentajes o número de vehículos)

Variable numérica:

- value: valor de unidades de la variable que aparece en ‘unit’.

6. Análisis y preprocesamiento de los datos

Global CO_2 and Greenhouse Gas Emissions

Importamos el *dataset* sobre el CO_2 y los gases de efecto invernadero, y le pasamos la lista de países que vamos a analizar:


```
# Create a list of 27 European countries to filter the chosen DataFrames:
europe_countries = ['Spain', 'France', 'Germany', 'Italy', 'Netherlands', 'Belgium', 'Austria', 'Sweden', 'Norway', 'Finland', 'Denmark', 'United Kingdom', 'Greece', 'Portugal', 'Ireland', 'Switzerla

# We import the data from the CO2 CSV file and filter it for European countries:
data_co2 = pd.read_csv('./data/owid-co2-data.csv')
data_co2_eu = data_co2[data_co2['country'].isin(europe_countries)]
```

Mostramos una parte:

	country	year	iso_code	population	gdp	cement_co2	cement_co2_per_capita	co2	co2_growth_abs	co2_growth_prct	...	share_global_other_co2	share_of_temperature_change_from_ghg	temperature_c
3441	Austria	1807	AUT	3124310.0	NaN	NaN	NaN	0.169	NaN	NaN	...	NaN	NaN	
3442	Austria	1819	AUT	3349015.0	NaN	NaN	NaN	0.253	NaN	50.000	...	NaN	NaN	
3443	Austria	1820	AUT	3367266.0	6.539229e+09	NaN	NaN	0.333	0.081	31.884	...	NaN	NaN	
3444	Austria	1821	AUT	3385004.0	NaN	NaN	NaN	0.359	0.026	7.692	...	NaN	NaN	
3445	Austria	1822	AUT	3402223.0	NaN	NaN	NaN	0.366	0.007	2.041	...	NaN	NaN	
3446	Austria	1823	AUT	3418916.0	NaN	NaN	NaN	0.348	-0.018	-5.000	...	NaN	NaN	
3447	Austria	1824	AUT	3435691.0	NaN	NaN	NaN	0.399	0.051	14.737	...	NaN	NaN	
3448	Austria	1825	AUT	3452549.0	NaN	NaN	NaN	0.403	0.004	0.917	...	NaN	NaN	
3449	Austria	1826	AUT	3469489.0	NaN	NaN	NaN	0.458	0.055	13.636	...	NaN	NaN	
3450	Austria	1827	AUT	3486512.0	NaN	NaN	NaN	0.476	0.018	4.000	...	NaN	NaN	

Luego pasamos la lista de variables escogidas, y vemos una descripción de las medidas de centro, desviación estándar, etc.:


```
# We filter the numerical data, only using specific variables for the correlation matrix:
data_co2_eu_num = data_co2_eu[['gdp', 'population', 'energy_per_gdp', 'consumption_co2', 'co2_per_capita', 'coal_co2_per_capita', 'co2_per_gdp', 'temperature_change_from_co2', 'nitrous_oxide_per_cap

# We show a summary of the data:
data_co2_eu_num.describe().round(3)
```

	gdp	population	energy_per_gdp	consumption_co2	co2_per_capita	coal_co2_per_capita	co2_per_gdp	temperature_change_from_co2	nitrous_oxide_per_capita	temperature_change_from_n2o	methane_p
count	3.381000e+03	4.839000e+03	1274.000	819.000	4539.000	4539.000	3320.000	4446.000	806.000	4446.000	
mean	2.360429e+11	1.400097e+07	2.082	185.412	4.125	2.179	0.482	0.002	0.767	0.000	
std	4.742450e+11	1.757739e+07	1.044	241.837	3.974	2.570	0.401	0.005	0.476	0.000	
min	1.540635e+09	3.768800e+05	0.593	11.874	0.000	0.000	0.000	0.000	0.269	0.000	
25%	2.286457e+10	3.130961e+06	1.354	45.606	0.554	0.320	0.212	0.000	0.459	0.000	
50%	7.021613e+10	5.918589e+06	1.745	82.442	3.012	1.208	0.368	0.000	0.596	0.000	
75%	2.097636e+11	1.639210e+07	2.545	216.674	6.928	3.096	0.610	0.001	0.956	0.000	
max	3.885961e+12	8.340856e+07	6.783	1207.438	25.534	17.077	3.378	0.041	2.996	0.002	

Vemos los valores nulos, los eliminamos y convertimos población en una variable de número entero:

```
# We check for missing values in the data:
data_co2_eu_num.isnull().sum()
```

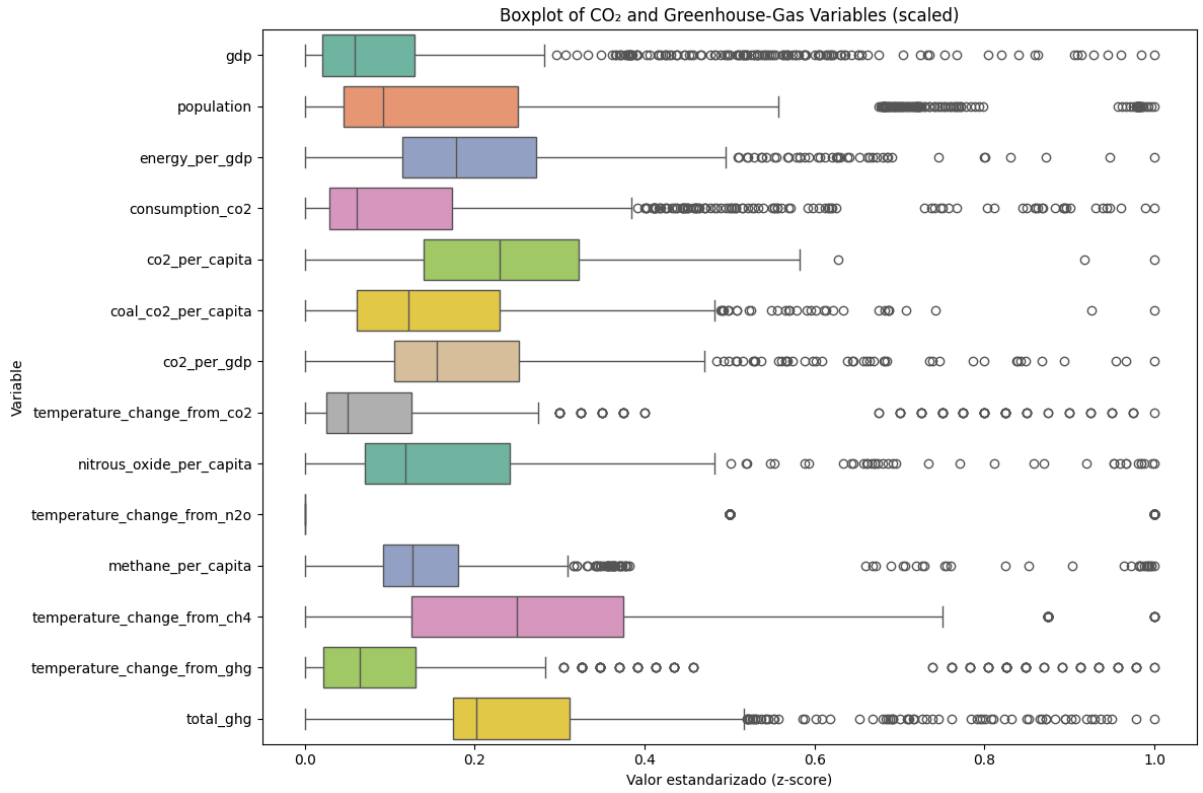
[1495] ✓ 0.0s

```
gdp                          1511
population                     53
energy_per_gdp               3618
consumption_co2              4073
co2_per_capita                353
coal_co2_per_capita           353
co2_per_gdp                  1572
temperature_change_from_co2   446
nitrous_oxide_per_capita     4086
temperature_change_from_n2o   446
methane_per_capita           4086
temperature_change_from_ch4   446
temperature_change_from_ghg   446
total_ghg                    4086
dtype: int64
```

```
# We delete the rows with missing values:
data_co2_eu_num = data_co2_eu_num.dropna()

# We convert `population` to int64:
data_co2_eu_num['population'] = data_co2_eu_num['population'].astype('int64')
```

Escalamos los datos con MinMaxScaler de Scikit-learn y mostramos un *boxplot* con Matplotlib y Seaborn:

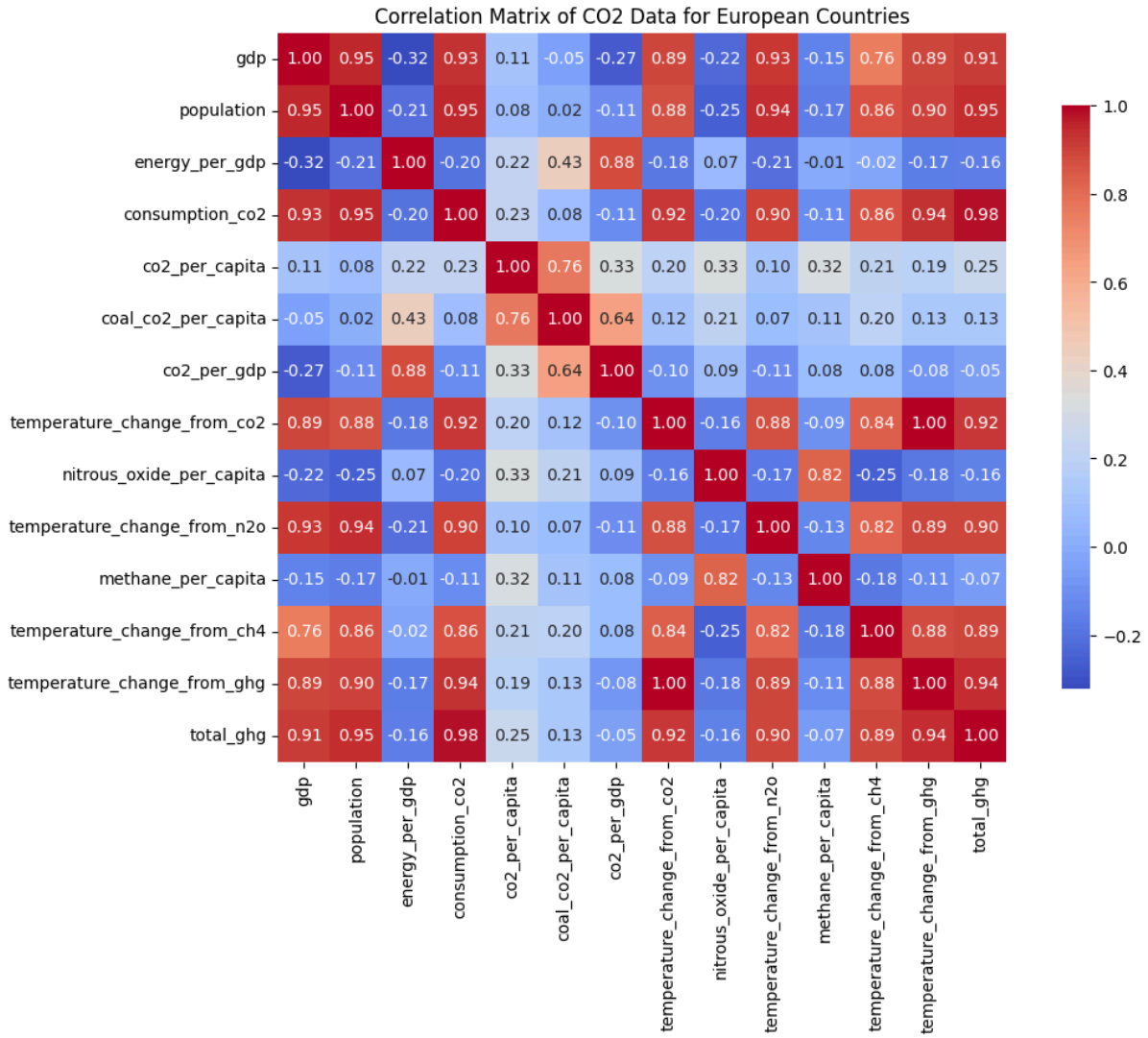


Observamos que la dispersión de los datos está principalmente a un nivel similar, pero hay muchos *outliers* en todas las variables.

Hacemos la matriz de correlación y la graficamos:

```
# We calculate the correlation matrix:
data_co2_eu_corr = data_co2_eu_num.corr()

# We plot the correlation matrix:
plt.figure(figsize=(12, 8))
sns.heatmap(data_co2_eu_corr, annot=True, fmt='.2f', cmap='coolwarm', square=True, cbar_kws={"shrink": .8})
plt.title('Correlation Matrix of CO2 Data for European Countries')
```



Vemos altas correlaciones positivas como el PIB con la población, el consumo de CO_2, los cambios de temperatura por CO_2, óxido nitroso, metano y gases de efecto invernadero. Y también otros bastante positivos como consumo de CO_2 y total de gases de efecto invernadero, cambio de temperatura por CO_2 con cambio de temperatura por gases de efecto invernadero (un 100% en este caso, lo cual puede indicar una alta colinealidad).

Y también se observan correlaciones negativas como PIB y energía por PIB y CO_2 por PIB. Población y óxido nitroso per cápita, etc.

De todas maneras, este es un examen preliminar de las variables, y está totalmente sujeto a la eliminación de las que puedan considerarse sobrantes.

Earth Temperature Data By Country (1743 - 2024)

Cargamos el *dataset* y mostramos una parte:

Data Preprocessing and exploration

```
data_temps_eu.head(10)
```

0.0s

	Year	Month	Adjusted Temperature	Country
3	1743	11	2.482	Austria
5	1743	11	7.106	Belgium
7	1743	11	5.928	Bulgaria
8	1743	11	7.225	Croatia
11	1743	11	0.727	Estonia
13	1743	11	-3.571	Finland
15	1743	11	10.203	France
16	1743	11	5.468	Germany
17	1743	11	10.806	Greece
19	1743	11	5.041	Hungary

Para quedarnos solo con el campo de año y eliminar los meses, generamos una nueva variable que calcula la media de las temperaturas ajustadas por año, y le llamamos ‘Average Temperature’:

```
# We create another dataframe out of this one, and what we need is to sum all the temperatures of each month, for each year and for each country, make a new column to make the average temperature for each year, and only leave
annual_avg_temps = (data_temps_eu.groupby(['Country', 'Year'], as_index=False)['Adjusted Temperature'].mean().rename(columns={'Adjusted Temperature': 'Average Temperature'}).copy())

# We make sure that it only contains the columns we need:
annual_avg_temps.head(20)
```

0.0s Python

	Country	Year	Average Temperature
0	Austria	1743	2.482000
1	Austria	1744	7.784750
2	Austria	1745	-0.977250
3	Austria	1746	NaN
4	Austria	1747	NaN
5	Austria	1748	NaN
6	Austria	1749	NaN
7	Austria	1750	6.983545
8	Austria	1751	7.185000
9	Austria	1752	1.821500
10	Austria	1753	6.138917
11	Austria	1754	5.907333
12	Austria	1755	5.671917
13	Austria	1756	6.463583
14	Austria	1757	6.097167
15	Austria	1758	5.363417
16	Austria	1759	6.363833
17	Austria	1760	6.379500
18	Austria	1761	6.572583
19	Austria	1762	6.080583

De este modo lo podremos unir al *dataset* original.

Comparamos las descripciones de la columna ‘Adjusted Temperature’, con la de ‘Average Temperature’, para ver las diferencias entre las medidas:

```
[1685]  data_temps_eu['Adjusted Temperature'].describe().round(3)
        ✓ 0.0s

...     count    107920.000
        mean          8.092
        std           9.179
        min         -36.830
        25%           2.352
        50%           8.872
        75%          15.108
        max          28.795
        Name: Adjusted Temperature, dtype: float64
```

```
[1686]  annual_avg_temps['Average Temperature'].describe().round(3)
        ✓ 0.0s

...     count    7146.000
        mean        7.961
        std         5.339
        min       -20.268
        25%         5.993
        50%         8.755
        75%        10.958
        max        17.698
        Name: Average Temperature, dtype: float64
```

Se pueden observar algunos grados de diferencia en alguna de las medidas, pero como nos interesan las medias para las visualizaciones, nos es útil haber hecho esa transformación.

Global Electric Vehicle Sales Data (2010-2024)

Importamos el último *dataset* y mostramos una parte:

```
data_ev_eu.head(10)
```
[1688] ✓ 0.0s

	region	category	parameter	mode	powertrain	year	unit	value
94	Austria	Historical	EV stock	Cars	BEV	2010	Vehicles	350.0000
95	Austria	Historical	EV stock share	Cars	EV	2010	percent	0.0079
96	Austria	Historical	EV stock share	Cars	EV	2011	percent	0.0220
97	Austria	Historical	EV charging points	EV	Publicly available fast	2011	charging points	0.1000
98	Austria	Historical	EV charging points	EV	Publicly available slow	2011	charging points	0.1000
99	Austria	Historical	EV stock	Cars	BEV	2011	Vehicles	990.0000
100	Austria	Historical	EV stock	Cars	BEV	2012	Vehicles	1400.0000
101	Austria	Historical	EV stock	Cars	PHEV	2012	Vehicles	140.0000
102	Austria	Historical	EV charging points	EV	Publicly available fast	2012	charging points	0.1000
103	Austria	Historical	EV charging points	EV	Publicly available slow	2012	charging points	0.1000

Mostramos la descripción de la columna ‘value’, siempre que ‘unit’ contenga ‘Vehicles’:

```
count       2575.0
mean       17195.8
std        75656.0
min            1.0
25%           25.0
50%          250.0
75%         3200.0
max      1500000.0
Name: value, dtype: float64
```

Creamos un nuevo *dataframe* que contenga la suma de los valores de ‘value’ que aparezcan cuando ‘unit’ tenga el valor ‘Vehicles’ en función de ‘region’ y ‘year’. De este modo, sumaremos todos los vehículos eléctricos, sean del tipo que sean, por país y por año:

```
# We create a new DataFrame where there is going to be a column called EV, and every cell of it will be a sum of all the values in the column 'value', of every year and for each country, where the unit is equal to 'Vehicles':
data_ev_eu_vehicles = data_ev_eu[data_ev_eu['unit'] == 'Vehicles'].copy()
data_ev_eu_vehicles = data_ev_eu_vehicles.groupby(['region', 'year'], as_index=False)['value'].sum().rename(columns={'value': 'EV'}).copy()
data_ev_eu_vehicles = data_ev_eu_vehicles.sort_values(by='year', ascending=False).reset_index(drop=True)

# We show an example of 15 rows where the region is equal to 'Spain':
data_ev_eu_vehicles[data_ev_eu_vehicles["region"] == "Spain"].head(15)
```

[1747] ✓ 0.0s Python

	region	year	EV
7	Spain	2023	519239
44	Spain	2022	327204
58	Spain	2021	235089
90	Spain	2020	140306
111	Spain	2019	71025
137	Spain	2018	46420
163	Spain	2017	27697
201	Spain	2016	15660
226	Spain	2015	9422
250	Spain	2014	5721
263	Spain	2013	2844
277	Spain	2012	1750
295	Spain	2011	890
309	Spain	2010	152

Mostramos la descripción de la nueva variable ‘EV’ que contiene esa suma de vehículos eléctricos:

```
data_ev_eu_vehicles['EV'].describe().round(3).reset_index()
```

[1748] ✓ 0.0s

	index	EV
0	count	313.000
1	mean	141466.923
2	std	378559.103
3	min	6.000
4	25%	1018.000
5	50%	8614.000
6	75%	96886.000
7	max	3331035.000

Unión de los *DataFrames*

Para acabar, uniremos los tres *dataframes* para enriquecer los datos.

Volvemos a crear el *dataframe* del CO_2 y los gases de efecto invernadero para escoger las variables categóricas y numéricas que nos quedamos de momento, y convertimos la población a entero:

Merging of the three DataFrames

```
# We create a new DataFrame where we only have the categorical and numerical variables we need:
data_co2_eu_merged = data_co2_eu[['country', 'year', 'gdp', 'population', 'energy_per_gdp', 'consumption_co2', 'co2_per_capita', 'coal_co2_per_capita', 'co2_per_gdp',  'temperature_change_from_co2', 'nitrous_oxide_per_capita'

# We convert 'population' to int64:
data_co2_eu_merged['population'] = data_co2_eu_merged['population'].astype('int64')
```

✓ 0.0s Python

```
data_co2_eu_merged[data_co2_eu_merged['country'] == 'Spain'].tail(30)
```

✓ 0.0s Python

	country	year	gdp	population	energy_per_gdp	consumption_co2	co2_per_capita	coal_co2_per_capita	co2_per_gdp	temperature_change_from_co2	nitrous_oxide_per_capita	temperature_change_from_n2o	methane_per_capita	ten
41300	Spain	1990	7.561250e+11	38889888	1.413	264.914	5.927	2.054	0.305	0.003	0.572	0.000	0.877	
41301	Spain	1991	7.873665e+11	38997064	1.405	270.629	6.160	2.113	0.305	0.003	0.551	0.000	0.876	
41302	Spain	1992	8.059446e+11	39202248	1.438	292.392	6.342	2.185	0.308	0.003	0.508	0.000	0.883	
41303	Spain	1993	7.992702e+11	39420540	1.403	272.007	6.089	2.092	0.300	0.004	0.517	0.000	0.868	
41304	Spain	1994	8.278065e+11	39623732	1.419	275.616	6.367	2.093	0.305	0.004	0.530	0.000	0.874	
41305	Spain	1995	8.655790e+11	39814568	1.409	287.613	6.682	2.157	0.307	0.004	0.520	0.000	0.880	
41306	Spain	1996	8.959665e+11	39996476	1.414	272.145	6.338	1.817	0.283	0.004	0.570	0.000	0.896	
41307	Spain	1997	9.373874e+11	40180048	1.426	277.698	6.616	1.875	0.284	0.004	0.560	0.000	0.937	
41308	Spain	1998	9.889724e+11	40362356	1.417	291.780	6.803	1.820	0.278	0.004	0.589	0.000	0.942	
41309	Spain	1999	1.044701e+12	40542232	1.382	315.556	7.352	2.060	0.285	0.004	0.585	0.001	0.950	
41310	Spain	2000	1.108745e+12	40741652	1.378	335.255	7.611	2.168	0.280	0.004	0.585	0.001	0.963	
41311	Spain	2001	1.164390e+12	40966448	1.372	335.740	7.611	1.980	0.268	0.005	0.571	0.001	0.972	
41312	Spain	2002	1.206113e+12	41477652	1.339	358.857	8.003	2.192	0.275	0.005	0.555	0.001	0.969	
41313	Spain	2003	1.254543e+12	42230268	1.361	358.650	7.964	1.996	0.268	0.005	0.569	0.001	0.951	
41314	Spain	2004	1.307501e+12	42959672	1.353	385.581	8.217	2.034	0.270	0.005	0.543	0.001	0.928	
41315	Spain	2005	1.365024e+12	43685372	1.305	411.236	8.432	2.009	0.270	0.005	0.503	0.001	0.906	
41316	Spain	2006	1.434224e+12	44422824	1.260	416.307	8.096	1.755	0.251	0.005	0.502	0.001	0.886	

Unimos el *dataframe* original al de las temperaturas, basándonos en el país y el año, y verificamos que la columna de la media de temperatura está ahí:

CO2 and Greenhouse Gases with Temperatures dataset merging

```
# We merge data_co2_eu_merged with annual_avg_temps by 'country' and 'year', and we rename the column 'Average Temperature' to 'avg_temp':
data_co2_temps_eu_merged = data_co2_eu_merged.merge(annual_avg_temps.rename(columns={'Average Temperature': 'avg_temp'}), left_on=['country', 'year'], right_on=['Country', 'Year'], how='left').drop(columns=['Country', 'Year']
data_co2_temps_eu_merged[data_co2_temps_eu_merged['country'] == 'Spain'].tail(30)
```

✓ 0.0s Python

	on_co2	co2_per_capita	coal_co2_per_capita	co2_per_gdp	temperature_change_from_co2	nitrous_oxide_per_capita	temperature_change_from_n2o	methane_per_capita	temperature_change_from_ch4	temperature_change_from_ghg	total_ghg	avg_temp
	264.914	5.927	2.054	0.305	0.003	0.572	0.000	0.877	0.001	0.005	232.94	14.871427
	270.629	6.160	2.113	0.305	0.003	0.551	0.000	0.876	0.001	0.005	239.03	14.096763
	292.392	6.342	2.185	0.308	0.003	0.508	0.000	0.883	0.001	0.005	248.88	14.030404
	272.007	6.089	2.092	0.300	0.004	0.517	0.000	0.868	0.002	0.006	234.07	13.600740
	275.616	6.367	2.093	0.305	0.004	0.530	0.000	0.874	0.002	0.006	246.98	14.903096
	287.613	6.682	2.157	0.307	0.004	0.520	0.000	0.880	0.002	0.006	260.70	15.357683
	272.145	6.338	1.817	0.283	0.004	0.570	0.000	0.896	0.002	0.006	254.09	14.365894
	277.698	6.616	1.875	0.284	0.004	0.560	0.000	0.937	0.002	0.006	275.54	15.331637
	291.780	6.803	1.820	0.278	0.004	0.589	0.000	0.942	0.002	0.006	284.99	14.775199
	315.556	7.352	2.060	0.285	0.004	0.585	0.001	0.950	0.002	0.006	309.58	14.598323
	335.255	7.611	2.168	0.280	0.004	0.585	0.001	0.963	0.002	0.007	326.90	14.737476
	335.740	7.611	1.980	0.268	0.005	0.571	0.001	0.972	0.002	0.007	348.60	14.852715
	358.857	8.003	2.192	0.275	0.005	0.555	0.001	0.969	0.002	0.007	365.30	14.896919
	358.650	7.964	1.996	0.268	0.005	0.569	0.001	0.951	0.002	0.007	374.30	15.218483
	385.581	8.217	2.034	0.270	0.005	0.543	0.001	0.928	0.002	0.007	390.41	14.605691
	411.236	8.432	2.009	0.270	0.005	0.503	0.001	0.906	0.002	0.008	405.06	14.519029
	416.307	8.096	1.755	0.251	0.005	0.502	0.001	0.886	0.002	0.008	399.06	15.357914
	428.442	8.120	1.870	0.245	0.005	0.505	0.001	0.885	0.002	0.008	415.34	14.416945
	393.270	7.307	1.291	0.220	0.006	0.448	0.001	0.837	0.002	0.008	382.03	14.379705
	345.547	6.395	0.972	0.199	0.006	0.437	0.001	0.866	0.002	0.008	345.93	15.162491
	328.190	6.075	0.781	0.189	0.006	0.445	0.001	0.829	0.002	0.008	331.00	14.366286

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS JUPYTER powershell

Para acabar, unimos este último *dataframe* con el de los vehículos eléctricos:

Y luego verificamos que, al final, está la columna del número de vehículos y podemos ver como, si filtramos por España como país, y porque los datos comienzan en 2010 en ese *dataset*, el recuento de vehículos eléctricos comienza ese año:

0.005	0.503	0.001	0.906	0.002	0.008	405.06	14.519029	NaN
0.005	0.502	0.001	0.886	0.002	0.008	399.06	15.357914	NaN
0.005	0.505	0.001	0.885	0.002	0.008	415.34	14.416945	NaN
0.006	0.448	0.001	0.837	0.002	0.008	382.03	14.379705	NaN
0.006	0.437	0.001	0.866	0.002	0.008	345.93	15.162491	NaN
0.006	0.445	0.001	0.829	0.002	0.008	331.00	14.366286	152.0
0.006	0.421	0.001	0.823	0.002	0.009	341.95	15.476688	890.0
0.006	0.407	0.001	0.807	0.002	0.009	335.52	14.811207	1750.0
0.006	0.420	0.001	0.798	0.002	0.009	309.12	14.736650	2844.0
0.006	0.439	0.001	0.797	0.002	0.009	307.85	15.597365	5721.0
0.006	0.442	0.001	0.831	0.002	0.009	318.39	15.732520	9422.0
0.007	0.437	0.001	0.828	0.002	0.009	302.82	15.504119	15660.0
0.007	0.445	0.001	0.835	0.002	0.009	319.43	15.986219	27697.0
0.007	0.449	0.001	0.837	0.002	0.010	312.92	15.063205	46420.0

Conclusión

En esta primera parte del trabajo se ha escogido el tema de la contaminación con los gases de efecto invernadero, y se ha enriquecido ese *dataset* con datos sobre las temperaturas globales registradas por país y por año, además de añadir la adopción de los vehículos eléctricos dentro de los mismos datos.

A pesar de que se han mantenido un gran número de variables, estas solamente son una parte de todas las que contenía el *dataset* original, pero como en el análisis exploratorio se han visto una buena cantidad de variables con una gran correlación, es probable que para la próxima parte se acaben eliminando unas cuantas para evitar una posible colinealidad que no nos aporte una información interesante al crear las visualizaciones.

Además, todo irá en función de si lo recopilado en este trabajo está evaluado como aceptable o si requiere cambios leves o más importantes.