Práctica II

1. Obtención de los datos

Funciones de utilidad

Importación de las librerías necesarias

In [213…
```
import seaborn as sns
import pandas as pd
import numpy as np
from scipy import stats
from scipy.stats import pearsonr
from scipy.stats import levene
import matplotlib.pyplot as plt
import statsmodels.api as sm
from scipy.stats import shapiro
from scipy.stats import bartlett
from scipy.stats import norm



from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import r2_score
from sklearn.metrics import mean_squared_error
```

Función **`export_dataframe`** que permite exportar el dataframe a fichero CSV

In [214…
```
def export_dataframe( df, file_name, directory ):
    file = 'C:\Python_Projects\TipologiaYCicloDeVidaDelDato\Data\%s/%s.csv' % (directory, file_name )
    #file = '/home/jovyan/work/%s/%s.csv' % (directory, file_name )
    df.to_csv(file )
```

In [215…
```
def read_dataframe( file_name, directory ):
    file = 'C:\Python_Projects\TipologiaYCicloDeVidaDelDato\Data\%s/%s.csv' % (directory, file_name )
    #file = '/home/jovyan/work/%s/%s.csv' % (directory, file_name )
    return pd.read_csv(file )
```

In [216…
```
def read_subdataframe( file_name ):
    return read_dataframe(
        file_name = file_name,
        directory = 'subdataset'
    ).iloc[ : , 1:]
```

DATASET I: Precio de Gas doméstico en € por kw/h

Obtención de los datos del subdataset de los precios del gas doméstico: `data_gas_prices_household_consumers.csv`

Importación del subdataframe

In [217…
```
data_gas_prices_household_consumers = read_subdataframe(
    file_name = 'data_gas_prices_household_consumers'
)
```

Columnas del dataset:

In [218…
```
display( data_gas_prices_household_consumers.dtypes )
```

```
country          object
country_name     object
2017            float64
2018            float64
2019            float64
2020            float64
2021            float64
dtype: object
```

Se muestran los 10 primeros valores:

In [219…
```
data_gas_prices_household_consumers.head( 10 )
```

Out[219]:

	country	country_name	2017	2018	2019	2020	2021
0	AT	Austria	0.0299	0.0304	0.0312	0.0308	0.0316
1	BA	Bosnia and Herzegovina	0.0240	0.0240	0.0249	0.0258	0.0251
2	BE	Belgium	0.0283	0.0288	0.0289	0.0252	0.0315
3	BG	Bulgaria	0.0170	0.0209	0.0240	0.0177	0.0331
4	CZ	Czechia	0.0360	0.0390	0.0455	0.0431	0.0448
5	DE	Germany (until 1990 former territory of the FRG)	NaN	NaN	0.0278	0.0292	0.0293
6	DK	Denmark	0.0234	0.0259	0.0209	0.0160	0.0415
7	EA	Euro area (EA11-1999, EA12-2001, EA13-2007, EA...	0.0295	0.0303	0.0319	0.0302	0.0315
8	EE	Estonia	0.0234	0.0239	0.0253	0.0240	0.0361
9	EL	Greece	NaN	0.0311	0.0338	0.0258	0.0444

DATASET II: Precio de Gas no doméstico en € por kw/h

Obtención de los datos del subdataset de los precios del gas no doméstico: `data_gas_prices_no_household_consumers.csv`

Importación del subdataframe

In [220…
```
data_gas_prices_no_household_consumers = read_subdataframe(
    file_name = 'data_gas_prices_no_household_consumers'
)
```

Columnas del dataset:

In [221…
```
display( data_gas_prices_no_household_consumers.dtypes )
```

```
country          object
country_name     object
2017            float64
2018            float64
2019            float64
2020            float64
2021            float64
dtype: object
```

Se muestran los 10 primeros valores:

In [222…
```
data_gas_prices_no_household_consumers.head( 10 )
```

Out[222]:

	country	country_name	2017	2018	2019	2020	2021
0	AT	Austria	NaN	NaN	0.0184	0.0168	0.0297
1	BA	Bosnia and Herzegovina	NaN	NaN	0.0257	0.0259	0.0248
2	BE	Belgium	NaN	NaN	0.0189	0.0148	0.0318
3	BG	Bulgaria	NaN	NaN	0.0213	0.0142	0.0299
4	CZ	Czechia	NaN	NaN	0.0226	0.0192	0.0259
5	DE	Germany (until 1990 former territory of the FRG)	NaN	NaN	0.0196	0.0171	0.0262
6	DK	Denmark	0.0194	0.0234	0.0178	0.0137	0.0448
7	EA	Euro area (EA11-1999, EA12-2001, EA13-2007, EA...	0.0220	0.0240	0.0211	0.0175	0.0278
8	EE	Estonia	NaN	NaN	0.0213	0.0155	0.0352
9	EL	Greece	NaN	NaN	0.0260	0.0165	0.0337

DATASET III: Precio de la electricidad doméstica para la franja de 2.500 a 4.999 kWh

Obtención de los datos del subdataset del precio de la electricidad doméstica para la franja de 2.500 a 4.999 kWh
`data_electricity_prices_household_consumers.csv`

Importación del subdataframe

In [223…
```
data_electricity_prices_household_consumers = read_subdataframe(
    file_name = 'data_electricity_prices_household_consumers'
)
```

Columnas del dataset:

In [224…
```
display( data_electricity_prices_household_consumers.dtypes )
```

```
country          object
country_name     object
2012-S2         float64
2013-S2         float64
2014-S2         float64
2015-S2         float64
2016-S2         float64
2017            float64
2018            float64
2019            float64
2020            float64
2021            float64
dtype: object
```

Se muestran los 10 primeros valores:

In [225…
```
data_electricity_prices_household_consumers.head(10)
```

Out[225]:

	country	country_name	2012-S2	2013-S2	2014-S2	2015-S2	2016-S2	2017	2018	2019	2020	2021
0	AL	Albania	NaN	NaN	NaN	NaN	NaN	0.0713	0.0759	0.0778	NaN	0.0781
1	AT	Austria	NaN	NaN	NaN	NaN	NaN	0.0613	0.0623	0.0687	0.0732	0.0745
2	BA	Bosnia and Herzegovina	NaN	NaN	NaN	NaN	NaN	0.0342	0.0338	0.0361	0.0365	NaN
3	BE	Belgium	NaN	NaN	NaN	NaN	NaN	0.0735	0.0808	0.0859	0.0786	0.0844
4	BG	Bulgaria	NaN	NaN	NaN	NaN	NaN	0.0575	0.0585	0.0558	0.0560	0.0608
5	CY	Cyprus	NaN	NaN	NaN	NaN	NaN	0.1036	0.1157	0.1241	0.1042	0.1094
6	CZ	Czechia	NaN	NaN	NaN	NaN	NaN	0.0541	0.0570	0.0690	0.0749	0.0979
7	DE	Germany (until 1990 former territory of the FRG)	NaN	NaN	NaN	NaN	NaN	0.0686	0.0622	0.0581	0.0574	0.0803
8	DK	Denmark	NaN	NaN	NaN	NaN	NaN	0.0388	0.0503	0.0539	0.0409	0.0747
9	EA	Euro area (EA11-1999, EA12-2001, EA13-2007, EA...	NaN	NaN	NaN	NaN	NaN	0.0760	0.0801	0.0727	0.0697	0.0898

DATASET IV: Precio de la electricidad no doméstica

OObtención de los datos del subdataset del precio de la electricidad no doméstica
`data_electricity_prices_no_household_consumers.csv`

Importación del subdataframe

In [226…
```
data_electricity_prices_no_household_consumers = read_subdataframe(
    file_name = 'data_electricity_prices_no_household_consumers'
)
```

Columnas del dataset:

In [227…
```
display( data_electricity_prices_no_household_consumers.dtypes )
```

```
country          object
country_name     object
2007-S2         float64
2008-S2         float64
2009-S2         float64
2010-S2         float64
2011-S2         float64
2012-S2         float64
2013-S2         float64
2014-S2         float64
2015-S2         float64
2016-S2         float64
2017            float64
2018            float64
2019            float64
2020            float64
2021            float64
dtype: object
```

Se muestran los 10 primeros valores:

In [228…
```
data_electricity_prices_no_household_consumers.head( 10 )
```

Out[228]:

	country	country_name	2007-S2	2008-S2	2009-S2	2010-S2	2011-S2	2012-S2	2013-S2	2014-S2	2015-S2	2016-S2	2017	2018	2019	2020	2021
0	AT	Austria	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.0598	0.0610	0.0654	0.0702	0.0723
1	BA	Bosnia and Herzegovina	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.0649	0.0621	0.0624	0.0648	NaN
2	BE	Belgium	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.0672	0.0624	0.0663	0.0745	0.0890
3	BG	Bulgaria	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.0817	0.0810	0.0764	0.0730	0.1075
4	CY	Cyprus	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.1187	0.1240	0.1271	0.1055	0.1136
5	CZ	Czechia	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.0580	0.0602	0.0721	0.0811	0.0848
6	DE	Germany (until 1990 former territory of the FRG)	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.0468	0.0612	0.0525	0.0651	0.0707
7	DK	Denmark	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.0433	0.0514	0.0517	0.0426	0.0898
8	EA	Euro area (EA11-1999, EA12-2001, EA13-2007, EA...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.0757	0.0837	0.0794	0.0780	0.0893
9	EE	Estonia	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.0406	0.0489	0.0516	0.0440	0.0850

2. Análisis inicial y procesamiento de los datos

Funciones de utilidad

Estimador utilizando la media `median_estimator`. Dentro de una columna del dataframe evalúa aquellos valores nulos y lo sustituye por la media de los valores que no lo son

In [229…
```
def median_estimator( df, column ) :
    median = df.loc[pd.notnull( df[column]), column].median()
    df[column].fillna(median,inplace=True)
    return df
```

Función `show_boxplot` que muestra el diagrama de caja de los valores de un dataframe

In [230…
```
def show_boxplot( df ):
    sns.set_theme( style = "whitegrid" )
    ax = sns.boxplot( data = df )
```

Función `init_outlier`. lializa el Outlier de una columna, inicializa a nulo el valor máximo de la columna

In [231…
```
def init_outlier(df, column):
    df.loc[
        df[column] == df[column].max(),
        column
    ] = np.nan
    return df
```

Datos de los costes del gas doméstico

Sustituimos NaN values por su media

In [232…
```
GasPricesHousehold=data_gas_prices_household_consumers

GasPricesHousehold = median_estimator( GasPricesHousehold, '2021' )
GasPricesHousehold = median_estimator( GasPricesHousehold, '2020' )
GasPricesHousehold = median_estimator( GasPricesHousehold, '2019' )
GasPricesHousehold = median_estimator( GasPricesHousehold, '2018' )
GasPricesHousehold = median_estimator( GasPricesHousehold, '2017' )
```

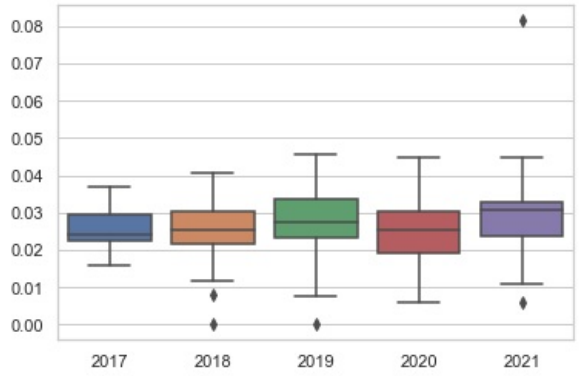
Se muestran los 10 primeros valores:

In [233…
```
display( GasPricesHousehold.head(10) )
```

	country	country_name	2017	2018	2019	2020	2021
0	AT	Austria	0.0299	0.0304	0.0312	0.0308	0.0316
1	BA	Bosnia and Herzegovina	0.0240	0.0240	0.0249	0.0258	0.0251
2	BE	Belgium	0.0283	0.0288	0.0289	0.0252	0.0315
3	BG	Bulgaria	0.0170	0.0209	0.0240	0.0177	0.0331
4	CZ	Czechia	0.0360	0.0390	0.0455	0.0431	0.0448
5	DE	Germany (until 1990 former territory of the FRG)	0.0243	0.0254	0.0278	0.0292	0.0293
6	DK	Denmark	0.0234	0.0259	0.0209	0.0160	0.0415
7	EA	Euro area (EA11-1999, EA12-2001, EA13-2007, EA...	0.0295	0.0303	0.0319	0.0302	0.0315
8	EE	Estonia	0.0234	0.0239	0.0253	0.0240	0.0361
9	EL	Greece	0.0243	0.0311	0.0338	0.0258	0.0444

Diagrama de caja para los diferentes años:

In [234…
```
show_boxplot( GasPricesHousehold )
```



Datos de los costes del gas para empresas

Sustituimos NaN values por su media

In [235…
```
GasPricesNoHousehold = data_gas_prices_no_household_consumers

GasPricesNoHousehold = median_estimator( GasPricesNoHousehold, '2021' )
GasPricesNoHousehold = median_estimator( GasPricesNoHousehold, '2020' )
GasPricesNoHousehold = median_estimator( GasPricesNoHousehold, '2019' )
GasPricesNoHousehold = median_estimator( GasPricesNoHousehold, '2018' )
GasPricesNoHousehold = median_estimator( GasPricesNoHousehold, '2017' )
```

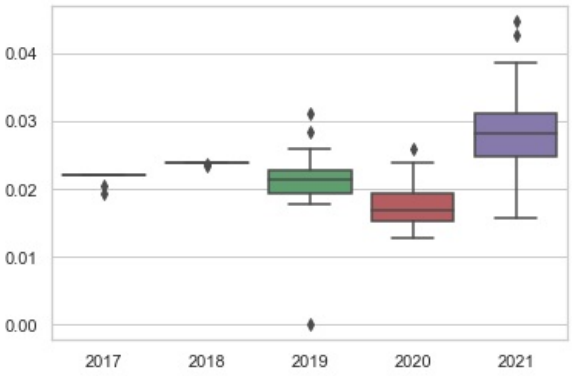
Se muestran los 10 primero valores:

In [236…
```
display( GasPricesNoHousehold.head(10) )
```

	country	country_name	2017	2018	2019	2020	2021
0	AT	Austria	0.0220	0.0240	0.0184	0.0168	0.0297
1	BA	Bosnia and Herzegovina	0.0220	0.0240	0.0257	0.0259	0.0248
2	BE	Belgium	0.0220	0.0240	0.0189	0.0148	0.0318
3	BG	Bulgaria	0.0220	0.0240	0.0213	0.0142	0.0299
4	CZ	Czechia	0.0220	0.0240	0.0226	0.0192	0.0259
5	DE	Germany (until 1990 former territory of the FRG)	0.0220	0.0240	0.0196	0.0171	0.0262
6	DK	Denmark	0.0194	0.0234	0.0178	0.0137	0.0448
7	EA	Euro area (EA11-1999, EA12-2001, EA13-2007, EA...	0.0220	0.0240	0.0211	0.0175	0.0278
8	EE	Estonia	0.0220	0.0240	0.0213	0.0155	0.0352
9	EL	Greece	0.0220	0.0240	0.0260	0.0165	0.0337

Diagrama de caja para los diferentes años:

In [237…
```
show_boxplot( GasPricesNoHousehold )
```



Datos de los costes de la electricidad doméstica

Se eliminan las columnas correspondientes a los valores semestrales de lo años desde el 2012 al 2016 que no contienen datos

In [238…
```
ElectPricesHouseholds = data_electricity_prices_household_consumers.drop(
    [
     '2012-S2',
     '2013-S2',
     '2014-S2',
     '2015-S2',
     '2016-S2'
    ], axis=1
)
```

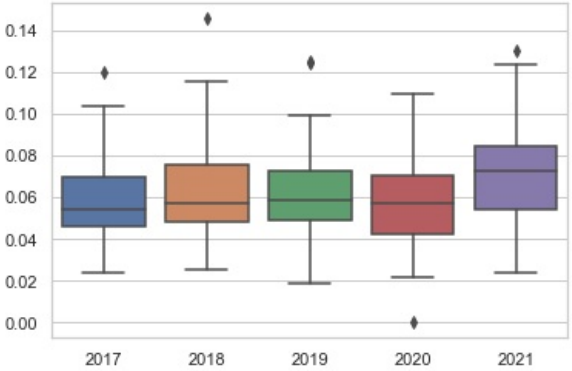
Muestra de los 10 primeros valores

In [239…
```
display( ElectPricesHouseholds.head( 10 ) )
```

	country	country_name	2017	2018	2019	2020	2021
0	AL	Albania	0.0713	0.0759	0.0778	NaN	0.0781
1	AT	Austria	0.0613	0.0623	0.0687	0.0732	0.0745
2	BA	Bosnia and Herzegovina	0.0342	0.0338	0.0361	0.0365	NaN
3	BE	Belgium	0.0735	0.0808	0.0859	0.0786	0.0844
4	BG	Bulgaria	0.0575	0.0585	0.0558	0.0560	0.0608
5	CY	Cyprus	0.1036	0.1157	0.1241	0.1042	0.1094
6	CZ	Czechia	0.0541	0.0570	0.0690	0.0749	0.0979
7	DE	Germany (until 1990 former territory of the FRG)	0.0686	0.0622	0.0581	0.0574	0.0803
8	DK	Denmark	0.0388	0.0503	0.0539	0.0409	0.0747
9	EA	Euro area (EA11-1999, EA12-2001, EA13-2007, EA...	0.0760	0.0801	0.0727	0.0697	0.0898

Se muestra el diagrama de caja

In [240…
```
show_boxplot( ElectPricesHouseholds )
```



Detectamos Outlier en el año 2021, inicializamos valor

In [241…
```
ElectPricesHouseholds = init_outlier(ElectPricesHouseholds, '2021')
```

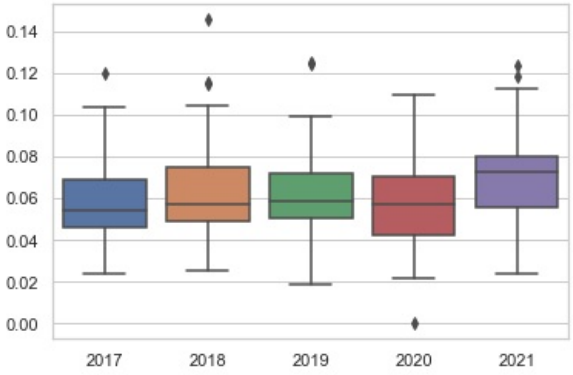
Estimamos valores nulos por la media

In [242…
```
ElectPricesHouseholds = median_estimator( ElectPricesHouseholds, '2021' )
ElectPricesHouseholds = median_estimator( ElectPricesHouseholds, '2020' )
ElectPricesHouseholds = median_estimator( ElectPricesHouseholds, '2019' )
```

```
ElectPricesHouseholds = median_estimator( ElectPricesHouseholds, '2018' )
ElectPricesHouseholds = median_estimator( ElectPricesHouseholds, '2017' )
```

Se vuelve a mostrar el diagrama de caja

In [243…
```
show_boxplot( ElectPricesHouseholds )
```



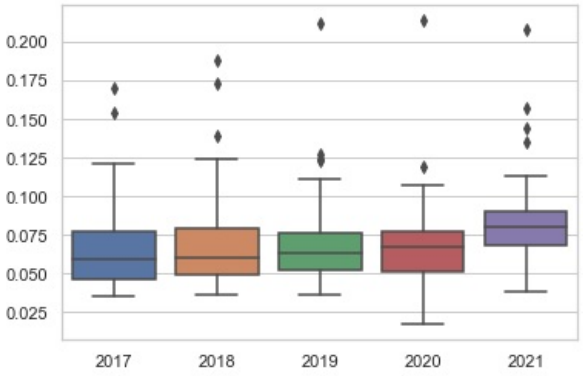
Datos de los costes de la electricidad para empresas

Se eliminan las columnas correspondientes a los valores semestrales de lo años desde el 2012 al 2016 que no contienen datos

In [244…
```
ElectPricesNoHouseholds = data_electricity_prices_no_household_consumers.drop(
    [
        '2007-S2',
        '2008-S2',
        '2009-S2',
        '2010-S2',
        '2011-S2',
        '2012-S2',
        '2013-S2',
        '2014-S2',
        '2015-S2',
        '2016-S2'
    ], axis=1
)
```

Se muestra el diagrama de caja

In [245…
```
show_boxplot( ElectPricesNoHouseholds )
```



Muestra de los 10 primeros valores

In [246…
```
display( ElectPricesNoHouseholds.head( 10 ) )
```

	country	country_name	2017	2018	2019	2020	2021
0	AT	Austria	0.0598	0.0610	0.0654	0.0702	0.0723
1	BA	Bosnia and Herzegovina	0.0649	0.0621	0.0624	0.0648	NaN
2	BE	Belgium	0.0672	0.0624	0.0663	0.0745	0.0890
3	BG	Bulgaria	0.0817	0.0810	0.0764	0.0730	0.1075
4	CY	Cyprus	0.1187	0.1240	0.1271	0.1055	0.1136
5	CZ	Czechia	0.0580	0.0602	0.0721	0.0811	0.0848
6	DE	Germany (until 1990 former territory of the FRG)	0.0468	0.0612	0.0525	0.0651	0.0707
7	DK	Denmark	0.0433	0.0514	0.0517	0.0426	0.0898
8	EA	Euro area (EA11-1999, EA12-2001, EA13-2007, EA...	0.0757	0.0837	0.0794	0.0780	0.0893
9	EE	Estonia	0.0406	0.0489	0.0516	0.0440	0.0850

Detectamos Outlier en los valores del 2021 y lo inicializamos

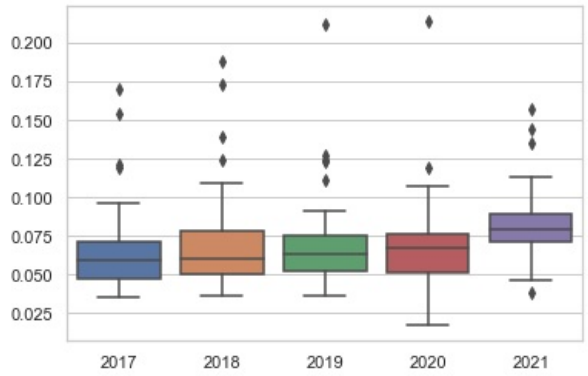
In [247…
```
ElectPricesHouseholds = init_outlier(ElectPricesNoHouseholds, '2021')
```

Se evaluan con el estimador de la media los valores nulos

In [248…
```
ElectPricesNoHouseholds = median_estimator( ElectPricesNoHouseholds, '2021' )
ElectPricesNoHouseholds = median_estimator( ElectPricesNoHouseholds, '2020' )
ElectPricesNoHouseholds = median_estimator( ElectPricesNoHouseholds, '2019' )
ElectPricesNoHouseholds = median_estimator( ElectPricesNoHouseholds, '2018' )
ElectPricesNoHouseholds = median_estimator( ElectPricesNoHouseholds, '2017' )
```

Se vuelve a mostrar el diagrama de caja

In [249…
```
show_boxplot( ElectPricesNoHouseholds )
```



Conclusión de análisis inicial de datos:

Descartamos continuar el análisis del "Gas" en el caso de los precios de los consumos de las empresas, ya que los datos obtenidos son una muestra demasiado pequeña, en concreto en los años 2017 y 2018 (ver dataset data_gas_prices_no_household_consumers.csv).

data_electricity_prices_household_consumers# Generación del dataset Final de trabajo

Generamos El Dataset final a exportar, con los datos posibles

In [250…
```
dElectCol = pd.merge(
    ElectPricesHouseholds,
    ElectPricesNoHouseholds.drop(['country_name'], axis=1),
    on='country',
    suffixes=('_ElectHouse', '_ElectNoHouse')
)

dGasCol = pd.merge(
    GasPricesHousehold.drop(['country_name'], axis=1),
    GasPricesNoHousehold.drop(['country_name'], axis=1),
    on='country',
    suffixes=('_GasHouse', '_GasNoHouse')
)

dEnergyCol = pd.merge(
    dElectCol,
    dGasCol,
    on='country'
)
```

Mostramos el dataset final a publicar:

In [251…
```
dEnergyCol.head()
```

Out[251]:

	country	country_name	2017_ElectHouse	2018_ElectHouse	2019_ElectHouse	2020_ElectHouse	2021_ElectHouse	2017_ElectNoHouse	20
0	AT	Austria	0.0598	0.0610	0.0654	0.0702	0.0723	0.0598	
1	BA	Bosnia and Herzegovina	0.0649	0.0621	0.0624	0.0648	0.0793	0.0649	
2	BE	Belgium	0.0672	0.0624	0.0663	0.0745	0.0890	0.0672	
3	BG	Bulgaria	0.0817	0.0810	0.0764	0.0730	0.1075	0.0817	
4	CZ	Czechia	0.0580	0.0602	0.0721	0.0811	0.0848	0.0580	

5 rows × 22 columns

Columnas del dataset a Publicar:

In [252…
```
display( dEnergyCol.dtypes )
```

```
country               object
country_name          object
2017_ElectHouse      float64
2018_ElectHouse      float64
2019_ElectHouse      float64
2020_ElectHouse      float64
2021_ElectHouse      float64
2017_ElectNoHouse    float64
2018_ElectNoHouse    float64
2019_ElectNoHouse    float64
2020_ElectNoHouse    float64
2021_ElectNoHouse    float64
2017_GasHouse        float64
2018_GasHouse        float64
2019_GasHouse        float64
2020_GasHouse        float64
2021_GasHouse        float64
2017_GasNoHouse      float64
2018_GasNoHouse      float64
2019_GasNoHouse      float64
2020_GasNoHouse      float64
2021_GasNoHouse      float64
dtype: object
```

Exportación dataset Final en formato CSV

In [253…
```
export_dataframe(
    df = dEnergyCol,
    file_name = 'energy_price_dataset',
    directory = 'dataset'
)
```

Dataset a analizar a partir del dataset publicado

De las conclusiones del anterior estudio vemos que no hay suficientes datos en los datos relativos al precio del gas de las empresas para poder hacer un análisis. Decimos entonces continuar sólo con los datos que hacen referencia a los precios del gas y de la electricidad relativos a entornos domésticos.

Costruimos un dataset filtrando solo estos datos, eliminando los datos relativos al precio del gas y a la electricidad de las empresas en el dataset original y también se eliminan los datos acumulados relativos a la Unión Europea:

In [254…
```
dEnergyHouseCol = dEnergyCol.loc[
                  : ,~dEnergyCol.columns.str.endswith('NoHouse')
                ].loc[
                  (dEnergyCol["country"] != "EU27_2020" )
                ].loc[
                  (dEnergyCol["country"] != "EA" )
                ]
```

Columnas del dataset a Analizar:

In [255…
```
display( dEnergyHouseCol.dtypes )
```

```
country             object
country_name        object
2017_ElectHouse    float64
2018_ElectHouse    float64
2019_ElectHouse    float64
2020_ElectHouse    float64
2021_ElectHouse    float64
2017_GasHouse      float64
2018_GasHouse      float64
2019_GasHouse      float64
2020_GasHouse      float64
2021_GasHouse      float64
dtype: object
```

Se presenta una muestra del dataset filtrando sólamente los datos domésticos

In [256…
```
display( dEnergyHouseCol.head() )
```

	country	country_name	2017_ElectHouse	2018_ElectHouse	2019_ElectHouse	2020_ElectHouse	2021_ElectHouse	2017_GasHouse	2018_G
0	AT	Austria	0.0598	0.0610	0.0654	0.0702	0.0723	0.0299	
1	BA	Bosnia and Herzegovina	0.0649	0.0621	0.0624	0.0648	0.0793	0.0240	
2	BE	Belgium	0.0672	0.0624	0.0663	0.0745	0.0890	0.0283	
3	BG	Bulgaria	0.0817	0.0810	0.0764	0.0730	0.1075	0.0170	
4	CZ	Czechia	0.0580	0.0602	0.0721	0.0811	0.0848	0.0360	

3. Análisis de los datos

Selección del grupo de datos

Teniendo por un lado un histórico de los precios de la electricidad y por otro los precios del gas por cada país. Se pretende hacer un estudio de la relación que existe entre ambos precios.

Para ello se procesa el dataset de los datos de precios de la energia doméstico para que cada registro tenga la información del país, del año y ambos precios

Crear el dataset de trabajo

Primero se crea una función `reduce_dataset` que permite añadir los precios de la electricidad y del gas. Cada registro tendrá la información del país, del año que se pasa como argumento, al que se refieren los precios, y las columnas del precio del gas y la electricidad.

In [257…
```
def reduce_dataset( original_df, year ):
    # Nombre de las columnas de la electricidad y gas del año pasado por argumento
    column_name_electricity =  '%s_ElectHouse' % year
    column_name_gas =  '%s_GasHouse' % year
    # Se obtiene las columnas relacionas con el país, el precio de la electricidad y el gas
    df =  original_df.loc[:,['country', column_name_electricity , column_name_gas]]
    # Se añade la columna del año
    df['Year'] = year
    # Se renombran las columnas de electricidad y gas por precio de electricidad y gas respectivamente
    return df.rename(
        columns= {
                    column_name_electricity: "ElectricityPrice",
                    column_name_gas        : "GasPrice"
                 }
    )
```

Se concatenan todos los años para crear el dataset de trabajo. También se resetea el índice del dataframe creado.

In [258…
```
df_work = pd.concat(
    [
        reduce_dataset( dEnergyHouseCol, 2017 ),
        reduce_dataset( dEnergyHouseCol, 2018 ),
        reduce_dataset( dEnergyHouseCol, 2019 ),
        reduce_dataset( dEnergyHouseCol, 2020 ),
        reduce_dataset( dEnergyHouseCol, 2021 )
    ]
)

df_work.reset_index(drop=True, inplace=True)
```

Se muestra los tipos de las columnas del dataset

In [259…
```
display( df_work.dtypes )
```

```
country              object
ElectricityPrice    float64
GasPrice            float64
Year                  int64
dtype: object
```

Se muestra un ejemplo de los datos del dataset de trabajo que consta de 160 registros

In [260…
```
display( df_work )
```

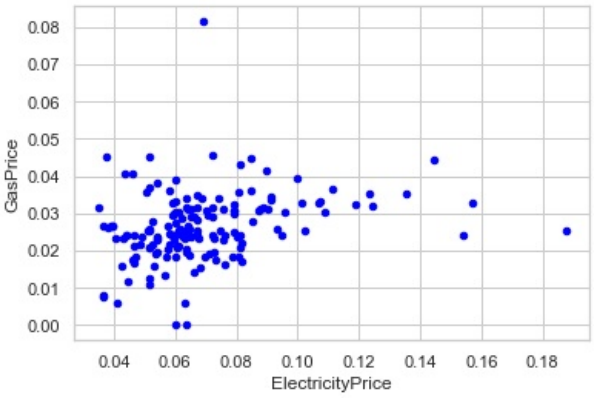
	country	ElectricityPrice	GasPrice	Year
0	AT	0.0598	0.0299	2017
1	BA	0.0649	0.0240	2017
2	BE	0.0672	0.0283	2017
3	BG	0.0817	0.0170	2017
4	CZ	0.0580	0.0360	2017
...	...	...	...	...
155	SI	0.0641	0.0264	2021
156	SK	0.0725	0.0195	2021
157	TR	0.0518	0.0110	2021
158	UA	0.0793	0.0306	2021
159	UK	0.0793	0.0306	2021

160 rows × 4 columns

Diagrama de dispersión entre los precios de la electricidad y del gas

Para comprobar la dependencia y la correlación entre los precios de la electricidad y el gas de manera gráfica, se muestra el diagrama de dispersión

In [261…
```
df_work.plot(kind='scatter', x = "ElectricityPrice", y ="GasPrice", color = "blue")

plt.show()
```



Valores de los precios de la electricidad

Se obtienen los valores del precio de la electricidad que se usarán en apartados posteriores:

In [262…
```
electricityPriceValues = df_work.loc[:, "ElectricityPrice"].to_numpy()

display( electricityPriceValues )
```

```
array([0.0598 , 0.0649 , 0.0672 , 0.0817 , 0.058  , 0.0468 , 0.0433 ,
       0.0406 , 0.081  , 0.1541 , 0.0641 , 0.05895, 0.0574 , 0.0536 ,
       0.0959 , 0.0949 , 0.0461 , 0.0581 , 0.0515 , 0.0675 , 0.0517 ,
       0.05895, 0.0629 , 0.0599 , 0.0468 , 0.0364 , 0.0348 , 0.044  ,
       0.0464 , 0.0533 , 0.05895, 0.0871 , 0.061  , 0.0621 , 0.0624 ,
       0.081  , 0.0602 , 0.0612 , 0.0514 , 0.0489 , 0.0876 , 0.1877 ,
       0.0636 , 0.0364 , 0.0613 , 0.0573 , 0.1091 , 0.1022 , 0.0488 ,
       0.058  , 0.054  , 0.0646 , 0.0505 , 0.0602 , 0.0609 , 0.0593 ,
       0.0539 , 0.0389 , 0.0434 , 0.051  , 0.0526 , 0.0446 , 0.0602 ,
       0.1015 , 0.0654 , 0.0624 , 0.0663 , 0.0764 , 0.0721 , 0.0525 ,
       0.0517 , 0.0516 , 0.0912 , 0.1242 , 0.0685 , 0.0363 , 0.0756 ,
       0.0695 , 0.1236 , 0.1112 , 0.0542 , 0.0615 , 0.0597 , 0.0609 ,
       0.0515 , 0.0636 , 0.0602 , 0.0674 , 0.0636 , 0.0395 , 0.0462 ,
       0.0577 , 0.0633 , 0.066  , 0.0636 , 0.0912 , 0.0702 , 0.0648 ,
       0.0745 , 0.073  , 0.0811 , 0.0651 , 0.0426 , 0.044  , 0.0933 ,
       0.0854 , 0.0757 , 0.0413 , 0.0728 , 0.068  , 0.1189 , 0.1067 ,
       0.0473 , 0.0607 , 0.0567 , 0.0637 , 0.054  , 0.067  , 0.067  ,
       0.0806 , 0.0711 , 0.0393 , 0.0373 , 0.0639 , 0.0721 , 0.0516 ,
       0.0599 , 0.067  , 0.0723 , 0.0793 , 0.089  , 0.1075 , 0.0848 ,
       0.0707 , 0.0898 , 0.085  , 0.1444 , 0.0901 , 0.0793 , 0.0634 ,
       0.0818 , 0.0762 , 0.1571 , 0.1353 , 0.0809 , 0.072  , 0.0787 ,
       0.0542 , 0.0996 , 0.0793 , 0.0651 , 0.0744 , 0.0801 , 0.0379 ,
       0.0692 , 0.0641 , 0.0725 , 0.0518 , 0.0793 , 0.0793 ])
```

Valores de los precios del gas

Se obtienen los valores del precio de la gas que se usarńa en apartados posteriores:

In [263…
```
gasPriceValues = df_work.loc[:, "GasPrice"].to_numpy()

display( gasPriceValues )
```

```
array([0.0299 , 0.024  , 0.0283 , 0.017  , 0.036  , 0.0243 , 0.0234 ,
       0.0234 , 0.0243 , 0.0243 , 0.031  , 0.0243 , 0.0206 , 0.019  ,
       0.0302 , 0.0243 , 0.0174 , 0.0216 , 0.0207 , 0.0234 , 0.0371 ,
       0.0293 , 0.0237 , 0.0333 , 0.0168 , 0.0264 , 0.0314 , 0.0243 ,
       0.0213 , 0.0159 , 0.0243 , 0.0308 , 0.0304 , 0.024  , 0.0288 ,
       0.0209 , 0.039  , 0.0254 , 0.0259 , 0.0239 , 0.0311 , 0.0254 ,
       0.0317 , 0.008  , 0.0212 , 0.0183 , 0.0304 , 0.0254 , 0.0216 ,
       0.0245 , 0.023  , 0.0186 , 0.0357 , 0.0305 , 0.0246 , 0.0326 ,
       0.0197 , 0.0266 , 0.0405 , 0.0255 , 0.0218 , 0.0118 , 0.     ,
       0.0329 , 0.0312 , 0.0249 , 0.0289 , 0.024  , 0.0455 , 0.0278 ,
       0.0209 , 0.0253 , 0.0338 , 0.0318 , 0.0342 , 0.0075 , 0.0229 ,
       0.0184 , 0.0351 , 0.0366 , 0.0238 , 0.0258 , 0.0227 , 0.021  ,
       0.0454 , 0.0339 , 0.0274 , 0.0347 , 0.0204 , 0.0267 , 0.0405 ,
       0.0267 , 0.0234 , 0.0142 , 0.     , 0.0346 , 0.0308 , 0.0258 ,
       0.0252 , 0.0177 , 0.0431 , 0.0292 , 0.016  , 0.024  , 0.0258 ,
       0.0277 , 0.031  , 0.0059 , 0.0234 , 0.0156 , 0.0325 , 0.0327 ,
       0.0183 , 0.0212 , 0.0133 , 0.0194 , 0.038  , 0.0316 , 0.0243 ,
       0.0357 , 0.0192 , 0.0265 , 0.045  , 0.0254 , 0.024  , 0.0126 ,
       0.0183 , 0.02555, 0.0316 , 0.0251 , 0.0315 , 0.0331 , 0.0448 ,
       0.0293 , 0.0415 , 0.0361 , 0.0444 , 0.0313 , 0.0325 , 0.0058 ,
       0.0221 , 0.0164 , 0.0328 , 0.0354 , 0.0182 , 0.0291 , 0.0184 ,
       0.0232 , 0.0396 , 0.03   , 0.0236 , 0.0341 , 0.0237 , 0.0263 ,
       0.0817 , 0.0264 , 0.0195 , 0.011  , 0.0306 , 0.0306 ])
```

4. Análisis de la normalidad y homogeneidad

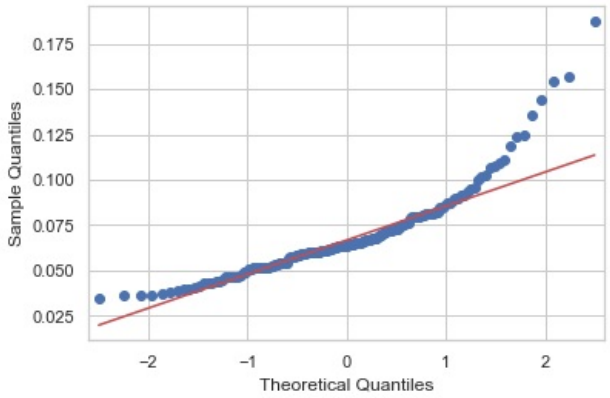
Normalidad de la muestra General

Análisis gráfico

Se crean los gráficos cuartil-cuartil para ver cuanto se aproximan a la normalidad ambas distribuciones

Precios de la electricidad

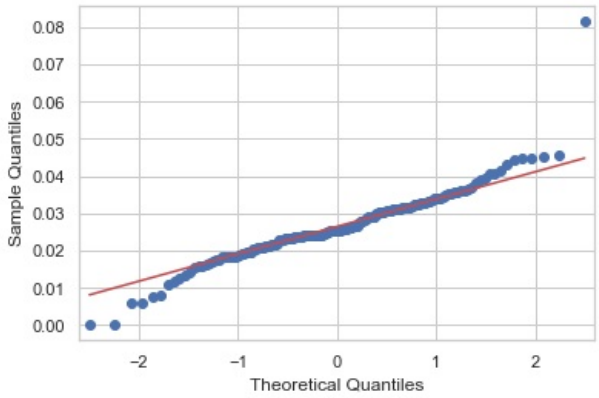
In [264…
```
sm.qqplot( electricityPriceValues, line='q' )

plt.show()
```



Se puede ver que existe valores centrales de la muestra que si se ajusta a una distribución normal.

Precios del gas

In [265…
```
fig2 = sm.qqplot( gasPriceValues, line='q' )
plt.show()
```



Exactamente que en el caso anterior.

Preparamos los datasets para trabajar en profundidad

In [266…
```
electricity = ElectPricesHouseholds[["country_name", "2017", "2018", "2019", "2020", "2021"]]
gas = GasPricesHousehold[["country_name", "2017", "2018", "2019", "2020", "2021"]]
```

In [267…
```
gas.loc[gas.country_name == "Germany (until 1990 former territory of the FRG)", "country_name"] = 'Germany'
electricity.loc[electricity.country_name == "Germany (until 1990 former territory of the FRG)", "country_name"]

gas.loc[gas.country_name == "European Union - 27 countries (from 2020)", "country_name"] = 'EU Country'
electricity.loc[electricity.country_name == "European Union - 27 countries (from 2020)", "country_name"] = 'EU

gas.loc[gas.country_name == "Euro area (EA11-1999, EA12-2001, EA13-2007, EA15-2008, EA16-2009, EA17-2011, EA18-
electricity.loc[electricity.country_name == "Euro area (EA11-1999, EA12-2001, EA13-2007, EA15-2008, EA16-2009,

gas.loc[gas.country_name == "Kosovo (under United Nations Security Council Resolution 1244/99)", "country_name"
electricity.loc[electricity.country_name == "Kosovo (under United Nations Security Council Resolution 1244/99)"
```

In [268…
```
gas_countries = gas.country_name.values
electric_countries = electricity.country_name.values
countries_both = [x for x in gas_countries if x in electric_countries]
electricity = electricity[electricity.country_name.isin(countries_both)]
gas = gas[gas.country_name.isin(countries_both)]
```

In [269…
```
pivoted_gas = pd.pivot_table(gas, columns = "country_name")
pivoted_electricity = pd.pivot_table(electricity, columns = "country_name")
```

4.1. Aplicación de pruebas estadísticas de forma generel y en profundidad

Pasamos a realizar los contrastes de hipótesis básicos para analizar el valor estadístico de calidad de la muestra

In [270…
```
def check_normality(df):
    groups = np.array([])
    for group in df.columns:
        groups = np.append(groups, df[group].values)
    stat, p = shapiro(groups[~np.isnan(groups)])
    print(f"Test shapiro, estadistico {stat} y pvalor {p}, H0 los datos siguen una distribucion normal\n")
    if p < 0.05:
        print(f"Probablemente los datos NO siguen una normal")
    else:
```

```
        print(f"Probablemente los datos SI siguen una normal")


def check_preconditions(df1, df2):
    print("-"*70)
    print("Checkeando normalidad para todo el dataframe")
    print("-"*70)
    print(f"Primer dataframe:")
    check_normality(df1)
    print("-"*70)
    print(f"Segundo dataframe:")
    check_normality(df2)
    print()
    print("-"*70)
    print("Checkeando homogeneidad de la varianza para los dos datasets")
    print("-"*70)
    check_homo_var(df1, df2)
```

In [271…
```
check_preconditions(pivoted_electricity, pivoted_gas)
```

```
----------------------------------------------------------------------
Checkeando normalidad para todo el dataframe
----------------------------------------------------------------------
Primer dataframe:
Test shapiro, estadistico 0.868183970451355 y pvalor 4.7169740335917254e-11, H0 los datos siguen una distribuci
on normal

Probablemente los datos NO siguen una normal
----------------------------------------------------------------------
Segundo dataframe:
Test shapiro, estadistico 0.9178404808044434 y pvalor 3.387682312450124e-08, H0 los datos siguen una distribuci
on normal

Probablemente los datos NO siguen una normal

----------------------------------------------------------------------
Checkeando homogeneidad de la varianza para los dos datasets
----------------------------------------------------------------------
Bartlett test para, H0 los datos tienen homogeneidad en las varianzas

Test bartlett, estadistico 130.8732383104749 y pvalor 2.6393361755798005e-30
Probablemente NO homogeneidad en las varianzas
```

En las muestras generales no encontramos ni normalidad ni correlación. Pasamos a analizarlo por cada muestra de cada país, pues por las gráficas iniciales, se ven subconjuntos con normalidad, probablemente sea por país. Además no podemos apoyarnos en el teorema central del limite, ya que tendriamos que asumir que el precio en los diferentes paises de gas y electricidad son independientes, asi que checkearemos pais por pais.

In [285…
```
def check_normality_each_country(df):
    for col in df.columns:
        stat, p = shapiro(df[col])
        print("-"*70)
        print(f"Test para {col}:")
        print(f"Test shapiro, estadistico {stat} y pvalor {p}, H0 los datos siguen una distribucion normal\n")
        if p < 0.05:
            print(f"Probablemente los datos NO siguen una normal\n")
        else:
            print(f"Probablemente los datos SI siguen una normal\n")
```

In [286…
```
check_normality_each_country(pivoted_gas)
```

```
----------------------------------------------------------------------
Test para Austria:
Test shapiro, estadistico 0.9899673461914062 y pvalor 0.9795799851417542, H0 los datos siguen una distribucion
normal

Probablemente los datos SI siguen una normal

----------------------------------------------------------------------
Test para Belgium:
Test shapiro, estadistico 0.9303292632102966 y pvalor 0.5986178517341614, H0 los datos siguen una distribucion
normal

Probablemente los datos SI siguen una normal

----------------------------------------------------------------------
Test para Bosnia and Herzegovina:
Test shapiro, estadistico 0.9009255766868591 y pvalor 0.41502219438552856, H0 los datos siguen una distribucion
normal

Probablemente los datos SI siguen una normal

----------------------------------------------------------------------
Test para Bulgaria:
Test shapiro, estadistico 0.8762538433074951 y pvalor 0.2926786243915558, H0 los datos siguen una distribucion
```

```
normal

Probablemente los datos SI siguen una normal

------------------------------------------------------------------------
Test para Croatia:
Test shapiro, estadistico 0.9603146910667419 y pvalor 0.8101732134819031, H0 los datos siguen una distribucion
normal

Probablemente los datos SI siguen una normal

------------------------------------------------------------------------
Test para Czechia:
Test shapiro, estadistico 0.9044991731643677 y pvalor 0.43524986505508423, H0 los datos siguen una distribucion
normal

Probablemente los datos SI siguen una normal

------------------------------------------------------------------------
Test para Denmark:
Test shapiro, estadistico 0.8861088752746582 y pvalor 0.3379155099391937, H0 los datos siguen una distribucion
normal

Probablemente los datos SI siguen una normal

------------------------------------------------------------------------
Test para EU Area:
Test shapiro, estadistico 0.9276749491691589 y pvalor 0.5806032419204712, H0 los datos siguen una distribucion
normal

Probablemente los datos SI siguen una normal

------------------------------------------------------------------------
Test para EU Country:
Test shapiro, estadistico 0.9457552433013916 y pvalor 0.7068579792976379, H0 los datos siguen una distribucion
normal

Probablemente los datos SI siguen una normal

------------------------------------------------------------------------
Test para Estonia:
Test shapiro, estadistico 0.6631652116775513 y pvalor 0.003786858869716525, H0 los datos siguen una distribucio
n normal

Probablemente los datos NO siguen una normal

------------------------------------------------------------------------
Test para France:
Test shapiro, estadistico 0.866613507270813 y pvalor 0.25298017263412476, H0 los datos siguen una distribucion
normal

Probablemente los datos SI siguen una normal

------------------------------------------------------------------------
Test para Georgia:
Test shapiro, estadistico 0.6589559316635132 y pvalor 0.0033876807428896427, H0 los datos siguen una distribuci
on normal

Probablemente los datos NO siguen una normal

------------------------------------------------------------------------
Test para Germany:
Test shapiro, estadistico 0.8805429339408875 y pvalor 0.31177860498428345, H0 los datos siguen una distribucion
normal

Probablemente los datos SI siguen una normal

------------------------------------------------------------------------
Test para Greece:
Test shapiro, estadistico 0.9151363968849182 y pvalor 0.49906301498413086, H0 los datos siguen una distribucion
normal

Probablemente los datos SI siguen una normal

------------------------------------------------------------------------
Test para Hungary:
Test shapiro, estadistico 0.8836076855659485 y pvalor 0.3259800672531128, H0 los datos siguen una distribucion
normal

Probablemente los datos SI siguen una normal

------------------------------------------------------------------------
Test para Ireland:
Test shapiro, estadistico 0.913966953754425 y pvalor 0.49179238080978394, H0 los datos siguen una distribucion
normal

Probablemente los datos SI siguen una normal

------------------------------------------------------------------------
```

```
Test para Italy:
Test shapiro, estadistico 0.8638687133789062 y pvalor 0.24247416853904724, H0 los datos siguen una distribucion
normal

Probablemente los datos SI siguen una normal

----------------------------------------------------------------------
Test para Latvia:
Test shapiro, estadistico 0.8821401596069336 y pvalor 0.31912195682525635, H0 los datos siguen una distribucion
normal

Probablemente los datos SI siguen una normal

----------------------------------------------------------------------
Test para Lithuania:
Test shapiro, estadistico 0.8642815947532654 y pvalor 0.24403248727321625, H0 los datos siguen una distribucion
normal

Probablemente los datos SI siguen una normal

----------------------------------------------------------------------
Test para Luxembourg:
Test shapiro, estadistico 0.9314118027687073 y pvalor 0.6060293912887573, H0 los datos siguen una distribucion
normal

Probablemente los datos SI siguen una normal

----------------------------------------------------------------------
Test para Moldova:
Test shapiro, estadistico 0.895110011100769 y pvalor 0.38345181941986084, H0 los datos siguen una distribucion
normal

Probablemente los datos SI siguen una normal

----------------------------------------------------------------------
Test para Netherlands:
Test shapiro, estadistico 0.9197801351547241 y pvalor 0.5285221934318542, H0 los datos siguen una distribucion
normal

Probablemente los datos SI siguen una normal

----------------------------------------------------------------------
Test para North Macedonia:
Test shapiro, estadistico 0.8768107295036316 y pvalor 0.2951079308986664, H0 los datos siguen una distribucion
normal

Probablemente los datos SI siguen una normal

----------------------------------------------------------------------
Test para Poland:
Test shapiro, estadistico 0.7781590819358826 y pvalor 0.0531456395983696, H0 los datos siguen una distribucion
normal

Probablemente los datos SI siguen una normal

----------------------------------------------------------------------
Test para Portugal:
Test shapiro, estadistico 0.9903190732002258 y pvalor 0.9808000922203064, H0 los datos siguen una distribucion
normal

Probablemente los datos SI siguen una normal

----------------------------------------------------------------------
Test para Romania:
Test shapiro, estadistico 0.9580806493759155 y pvalor 0.7945586442947388, H0 los datos siguen una distribucion
normal

Probablemente los datos SI siguen una normal

----------------------------------------------------------------------
Test para Serbia:
Test shapiro, estadistico 0.9867621660232544 y pvalor 0.9671739339828491, H0 los datos siguen una distribucion
normal

Probablemente los datos SI siguen una normal

----------------------------------------------------------------------
Test para Slovakia:
Test shapiro, estadistico 0.9602929949760437 y pvalor 0.8100219368934631, H0 los datos siguen una distribucion
normal

Probablemente los datos SI siguen una normal

----------------------------------------------------------------------
Test para Slovenia:
Test shapiro, estadistico 0.9442073702812195 y pvalor 0.6958163976669312, H0 los datos siguen una distribucion
normal

Probablemente los datos SI siguen una normal
```

```
------------------------------------------------------------------------
Test para Spain:
Test shapiro, estadistico 0.8963978290557861 y pvalor 0.3902978301048279, H0 los datos siguen una distribucion
normal

Probablemente los datos SI siguen una normal

------------------------------------------------------------------------
Test para Sweden:
Test shapiro, estadistico 0.7774146795272827 y pvalor 0.05236475169658661, H0 los datos siguen una distribucion
normal

Probablemente los datos SI siguen una normal

------------------------------------------------------------------------
Test para Turkey:
Test shapiro, estadistico 0.9555109143257141 y pvalor 0.7764301300048828, H0 los datos siguen una distribucion
normal

Probablemente los datos SI siguen una normal

------------------------------------------------------------------------
Test para Ukraine:
Test shapiro, estadistico 0.8690097332000732 y pvalor 0.26243698596954346, H0 los datos siguen una distribucion
normal

Probablemente los datos SI siguen una normal

------------------------------------------------------------------------
Test para United Kingdom:
Test shapiro, estadistico 0.9300695657730103 y pvalor 0.5968450903892517, H0 los datos siguen una distribucion
normal

Probablemente los datos SI siguen una normal
```

Efectivamente tal y como nos hizo sospechar las gráficas inciales si encontramos países cuyos precios se ajustan a la normalidad.

5.2 Estudio de la correlación

In [274…

```python
def check_homo_var(df1, df2):
    group1 = np.array([])
    for group in df1.columns:
        group1 = np.append(group1, df1[group].values)
    group2 = np.array([])
    for group in df2.columns:
        group2 = np.append(group2, df2[group].values)
    stat, p = bartlett(group1[~np.isnan(group1)], group2[~np.isnan(group2)])

    print(f"Bartlett test para, H0 los datos tienen homogeneidad en las varianzas\n")
    print(f"Test bartlett, estadistico {stat} y pvalor {p}")
    if p < 0.05:
        print(f"Probablemente NO homogeneidad en las varianzas\n")
    else:
        print(f"Probablemente SI homogeneidad en las varianzas\n")


def check_correlation(df1, df2):
    group1 = np.array([])
    for group in df1.columns:
        group1 = np.append(group1, df1[group].values)
    group2 = np.array([])
    for group in df2.columns:
        group2 = np.append(group2, df2[group].values)

    indxs1 = np.isnan(group1)
    indxs2 = np.isnan(group2)
    indxs = indxs1 | indxs2 # indices que tienen nan en alguno de los dos df

    group1 = group1[~indxs]
    group2 = group2[~indxs]

    stat, p = pearsonr(group1, group2)
    print(f"Test pearson, estadistico {stat} y pvalor {p}, H0 no estan correlacionadas\n")
    if p < 0.05:
        print(f"Probablemente NO estan correlacionados\n")
    else:
        print(f"Probablemente SI estan correlacionados\n")
```

In [275…

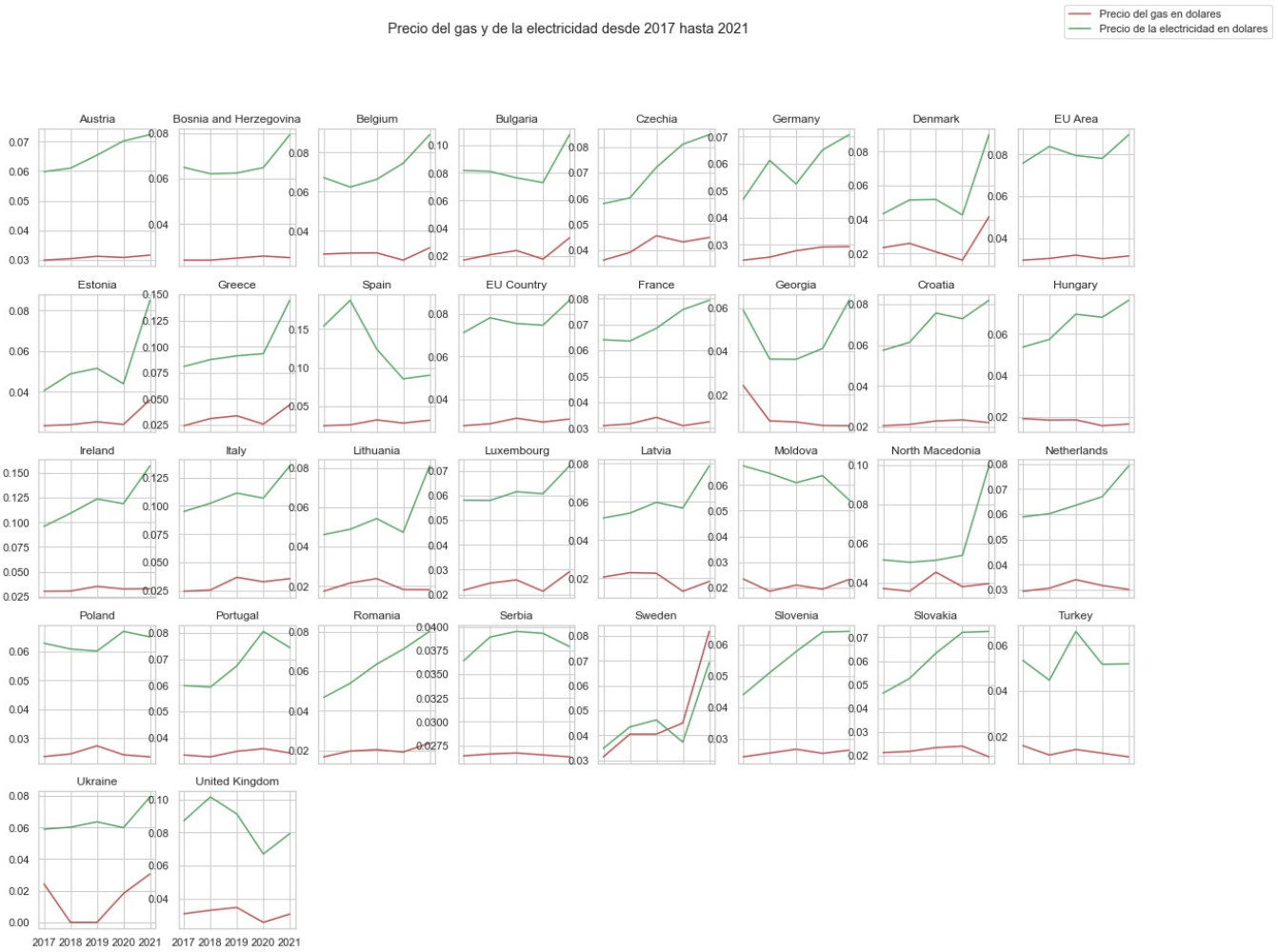
```python
check_homo_var(pivoted_electricity, pivoted_gas)
check_correlation(pivoted_electricity, pivoted_gas)
```

```
Bartlett test para, H0 los datos tienen homogeneidad en las varianzas

Test bartlett, estadistico 130.8732383104749 y pvalor 2.6393361755798005e-30
Probablemente NO homogeneidad en las varianzas

Test pearson, estadistico 0.282540907419713 y pvalor 0.00018917062353675837, H0 no estan correlacionadas

Probablemente NO estan correlacionados
```

Efectivamente, como en el caso anterior, no nos da correlación general de la muestra y además encontramos heterocedasticidad (Varianzas no constantes en la muestra, esto es un problema), pasamos a realizar por país.

In [276…
```python
def plot_countries(gas, electricity, names):
    fig, axs = plt.subplots(5, 8, figsize=(20,15), sharex=True)
    k = 0
    for i in range(5):
        for j in range(8):
            if k == len(names):
                fig.delaxes(axs[i][j])
                continue
            axs[i][j].plot(pivoted_gas.index, pivoted_gas[names[k]], color = 'r')
            axs[i][j].plot(pivoted_electricity.index, pivoted_electricity[names[k]], color = 'g')
            axs[i][j].set_title(f"{names[k]}")
            k += 1
    fig.legend(['Precio del gas en dolares', 'Precio de la electricidad en dolares'], loc='upper right')
    fig.suptitle("Precio del gas y de la electricidad desde 2017 hasta 2021")
    plt.show()
```

In [277…
```python
plot_countries(pivoted_gas, pivoted_electricity, countries_both)
```



Aparentemente vemos correlación directa e inversa en los conjuntos por países, pasamos a chequear.

In [278…
```python
def check_homo_var_each_country(df1, df2):
    group1 = np.array([])
    group2 = np.array([])
    for group in df1.columns:
        group1 = np.append(group1, df1[group].values)
        group2 = np.append(group2, df2[group].values)
        stat, p = bartlett(group1[~np.isnan(group1)], group2[~np.isnan(group2)])
        print("-"*70)
        print(f"Bartlett test para {group}, H0 los datos tienen homogeneidad en las varianzas\n")
        print(f"Estadistico {stat} y pvalor {p}")
        if p < 0.05:
            print(f"probablemente NO homogeneidad en las varianzas")
```

```
        else:
            print(f"probablemente SI homogeneidad en las varianzas")
        print("-"*70)
```

In [279…

```
check_homo_var_each_country(pivoted_gas, pivoted_electricity)
```

```
----------------------------------------------------------------------
Bartlett test para Austria, H0 los datos tienen homogeneidad en las varianzas

Estadistico 10.193735212469518 y pvalor 0.0014091858076003634
probablemente NO homogeneidad en las varianzas
----------------------------------------------------------------------
----------------------------------------------------------------------
Bartlett test para Belgium, H0 los datos tienen homogeneidad en las varianzas

Estadistico 14.191634787699723 y pvalor 0.00016510287297749078
probablemente NO homogeneidad en las varianzas
----------------------------------------------------------------------
----------------------------------------------------------------------
Bartlett test para Bosnia and Herzegovina, H0 los datos tienen homogeneidad en las varianzas

Estadistico 11.896558954264867 y pvalor 0.0005623773997189065
probablemente NO homogeneidad en las varianzas
----------------------------------------------------------------------
----------------------------------------------------------------------
Bartlett test para Bulgaria, H0 los datos tienen homogeneidad en las varianzas

Estadistico 14.074119370282013 y pvalor 0.00017574544528976732
probablemente NO homogeneidad en las varianzas
----------------------------------------------------------------------
----------------------------------------------------------------------
Bartlett test para Croatia, H0 los datos tienen homogeneidad en las varianzas

Estadistico 17.06256083524616 y pvalor 3.616834643278469e-05
probablemente NO homogeneidad en las varianzas
----------------------------------------------------------------------
----------------------------------------------------------------------
Bartlett test para Czechia, H0 los datos tienen homogeneidad en las varianzas

Estadistico 4.3748755932672685 y pvalor 0.03647249268879289
probablemente NO homogeneidad en las varianzas
----------------------------------------------------------------------
----------------------------------------------------------------------
Bartlett test para Denmark, H0 los datos tienen homogeneidad en las varianzas

Estadistico 9.901057401760188 y pvalor 0.0016518383204098415
probablemente NO homogeneidad en las varianzas
----------------------------------------------------------------------
----------------------------------------------------------------------
Bartlett test para EU Area, H0 los datos tienen homogeneidad en las varianzas

Estadistico 13.30503294380453 y pvalor 0.0002646946406197282
probablemente NO homogeneidad en las varianzas
----------------------------------------------------------------------
----------------------------------------------------------------------
Bartlett test para EU Country, H0 los datos tienen homogeneidad en las varianzas

Estadistico 15.919645386923577 y pvalor 6.608916704349897e-05
probablemente NO homogeneidad en las varianzas
----------------------------------------------------------------------
----------------------------------------------------------------------
Bartlett test para Estonia, H0 los datos tienen homogeneidad en las varianzas

Estadistico 24.97889757879068 y pvalor 5.796123735133159e-07
probablemente NO homogeneidad en las varianzas
----------------------------------------------------------------------
----------------------------------------------------------------------
Bartlett test para France, H0 los datos tienen homogeneidad en las varianzas

Estadistico 27.108414125998166 y pvalor 1.9235887416992228e-07
probablemente NO homogeneidad en las varianzas
----------------------------------------------------------------------
----------------------------------------------------------------------
Bartlett test para Georgia, H0 los datos tienen homogeneidad en las varianzas

Estadistico 18.88599928535714 y pvalor 1.3876739286586268e-05
probablemente NO homogeneidad en las varianzas
----------------------------------------------------------------------
----------------------------------------------------------------------
Bartlett test para Germany, H0 los datos tienen homogeneidad en las varianzas

Estadistico 22.1374923037241 y pvalor 2.538041664642065e-06
probablemente NO homogeneidad en las varianzas
----------------------------------------------------------------------
----------------------------------------------------------------------
Bartlett test para Greece, H0 los datos tienen homogeneidad en las varianzas

Estadistico 38.14211908163463 y pvalor 6.577486665513411e-10
```

```
probablemente NO homogeneidad en las varianzas
------------------------------------------------------------------------
------------------------------------------------------------------------
Bartlett test para Hungary, H0 los datos tienen homogeneidad en las varianzas

Estadistico 37.21546165417484 y pvalor 1.0577181358193716e-09
probablemente NO homogeneidad en las varianzas
------------------------------------------------------------------------
------------------------------------------------------------------------
Bartlett test para Ireland, H0 los datos tienen homogeneidad en las varianzas

Estadistico 66.23436486494516 y pvalor 4.003710435951206e-16
probablemente NO homogeneidad en las varianzas
------------------------------------------------------------------------
------------------------------------------------------------------------
Bartlett test para Italy, H0 los datos tienen homogeneidad en las varianzas

Estadistico 80.23288164583572 y pvalor 3.327830384087931e-19
probablemente NO homogeneidad en las varianzas
------------------------------------------------------------------------
------------------------------------------------------------------------
Bartlett test para Latvia, H0 los datos tienen homogeneidad en las varianzas

Estadistico 82.93412711186784 y pvalor 8.48338378852218e-20
probablemente NO homogeneidad en las varianzas
------------------------------------------------------------------------
------------------------------------------------------------------------
Bartlett test para Lithuania, H0 los datos tienen homogeneidad en las varianzas

Estadistico 88.10253164733152 y pvalor 6.2148034751039825e-21
probablemente NO homogeneidad en las varianzas
------------------------------------------------------------------------
------------------------------------------------------------------------
Bartlett test para Luxembourg, H0 los datos tienen homogeneidad en las varianzas

Estadistico 93.10026708188153 y pvalor 4.971102930991774e-22
probablemente NO homogeneidad en las varianzas
------------------------------------------------------------------------
------------------------------------------------------------------------
Bartlett test para Moldova, H0 los datos tienen homogeneidad en las varianzas

Estadistico 96.75810527672509 y pvalor 7.833625135454271e-23
probablemente NO homogeneidad en las varianzas
------------------------------------------------------------------------
------------------------------------------------------------------------
Bartlett test para Netherlands, H0 los datos tienen homogeneidad en las varianzas

Estadistico 100.44505941704239 y pvalor 1.2172708541070965e-23
probablemente NO homogeneidad en las varianzas
------------------------------------------------------------------------
------------------------------------------------------------------------
Bartlett test para North Macedonia, H0 los datos tienen homogeneidad en las varianzas

Estadistico 98.03947854447468 y pvalor 4.1012397504185996e-23
probablemente NO homogeneidad en las varianzas
------------------------------------------------------------------------
------------------------------------------------------------------------
Bartlett test para Poland, H0 los datos tienen homogeneidad en las varianzas

Estadistico 102.51138960616831 y pvalor 4.288947192572711e-24
probablemente NO homogeneidad en las varianzas
------------------------------------------------------------------------
------------------------------------------------------------------------
Bartlett test para Portugal, H0 los datos tienen homogeneidad en las varianzas

Estadistico 104.1702519072259 y pvalor 1.856572637140518e-24
probablemente NO homogeneidad en las varianzas
------------------------------------------------------------------------
------------------------------------------------------------------------
Bartlett test para Romania, H0 los datos tienen homogeneidad en las varianzas

Estadistico 106.64895618609656 y pvalor 5.314400727318254e-25
probablemente NO homogeneidad en las varianzas
------------------------------------------------------------------------
------------------------------------------------------------------------
Bartlett test para Serbia, H0 los datos tienen homogeneidad en las varianzas

Estadistico 119.76638831155694 y pvalor 7.116695305647977e-28
probablemente NO homogeneidad en las varianzas
------------------------------------------------------------------------
------------------------------------------------------------------------
Bartlett test para Slovakia, H0 los datos tienen homogeneidad en las varianzas

Estadistico 124.02226128415634 y pvalor 8.33037531088952e-29
probablemente NO homogeneidad en las varianzas
------------------------------------------------------------------------
------------------------------------------------------------------------
Bartlett test para Slovenia, H0 los datos tienen homogeneidad en las varianzas
```

```
Estadistico 130.3992060007176 y pvalor 3.351242410206087e-30
probablemente NO homogeneidad en las varianzas
------------------------------------------------------------------------
------------------------------------------------------------------------
Bartlett test para Spain, H0 los datos tienen homogeneidad en las varianzas

Estadistico 175.12562654168767 y pvalor 5.620476388605642e-40
probablemente NO homogeneidad en las varianzas
------------------------------------------------------------------------
------------------------------------------------------------------------
Bartlett test para Sweden, H0 los datos tienen homogeneidad en las varianzas

Estadistico 138.3844729391353 y pvalor 6.004907815608491e-32
probablemente NO homogeneidad en las varianzas
------------------------------------------------------------------------
------------------------------------------------------------------------
Bartlett test para Turkey, H0 los datos tienen homogeneidad en las varianzas

Estadistico 134.82911528249795 y pvalor 3.598441199985169e-31
probablemente NO homogeneidad en las varianzas
------------------------------------------------------------------------
------------------------------------------------------------------------
Bartlett test para Ukraine, H0 los datos tienen homogeneidad en las varianzas

Estadistico 125.62764519962633 y pvalor 3.7094553526048056e-29
probablemente NO homogeneidad en las varianzas
------------------------------------------------------------------------
------------------------------------------------------------------------
Bartlett test para United Kingdom, H0 los datos tienen homogeneidad en las varianzas

Estadistico 130.8732383104749 y pvalor 2.6393361755798005e-30
probablemente NO homogeneidad en las varianzas
------------------------------------------------------------------------
```

Efectivamente, volvemos a obtener heterocedasticidad, diferencia de varianzas en las muestras, esto es un problema, deberíamos trabajar las muestras más a profundidad para intentar solventarla, pero se sale de scope. Vamos, aún así, a estudiar la correlación.

In [280…
```python
def check_correlation_each_country(df1, df2):
    group1 = np.array([])
    group2 = np.array([])
    for group in df1.columns:
        group1 = np.append(group1, df1[group].values)
        group2 = np.append(group2, df2[group].values)

        indxs1 = np.isnan(group1)
        indxs2 = np.isnan(group2)
        indxs = indxs1 | indxs2 # indices que tienen nan en alguno de los dos df

        group1 = group1[~indxs]
        group2 = group2[~indxs]

        stat, p = pearsonr(group1, group2)
        print(f"Test pearson para {group}, estadistico {stat} y pvalor {p}, H0 no estan correlacionadas\n")
        if p < 0.05:
            print(f"Probablemente NO estan correlacionados")
        else:
            print(f"Probablemente SI esten correlacionados")
```

In [281…
```python
check_correlation_each_country(pivoted_electricity, pivoted_gas)
```

```
Test pearson para Austria, estadistico 0.845371068323818 y pvalor 0.07127402761509678, H0 no estan correlaciona
das

Probablemente SI esten correlacionados
Test pearson para Belgium, estadistico 0.10983703213365414 y pvalor 0.7626092116466721, H0 no estan correlacion
adas

Probablemente SI esten correlacionados
Test pearson para Bosnia and Herzegovina, estadistico 0.17972750265742504 y pvalor 0.5215576763339973, H0 no es
tan correlacionadas

Probablemente SI esten correlacionados
Test pearson para Bulgaria, estadistico 0.03969110130797798 y pvalor 0.8680472606214779, H0 no estan correlacio
nadas

Probablemente SI esten correlacionados
Test pearson para Croatia, estadistico 0.09655801836832324 y pvalor 0.6461275439071115, H0 no estan correlacion
adas

Probablemente SI esten correlacionados
Test pearson para Czechia, estadistico 0.1070330077542502 y pvalor 0.5734656580929945, H0 no estan correlaciona
das

Probablemente SI esten correlacionados
Test pearson para Denmark, estadistico 0.32926913448388684 y pvalor 0.05342336639497488, H0 no estan correlacio
```

```
nadas

Probablemente SI esten correlacionados
Test pearson para EU Area, estadistico 0.34884885671723004 y pvalor 0.027371765263910254, H0 no estan correlaci
onadas

Probablemente NO estan correlacionados
Test pearson para EU Country, estadistico 0.352600600833549 y pvalor 0.017519783852434266, H0 no estan correlac
ionadas

Probablemente NO estan correlacionados
Test pearson para Estonia, estadistico 0.40591687758966166 y pvalor 0.003447589877845323, H0 no estan correlaci
onadas

Probablemente NO estan correlacionados
Test pearson para France, estadistico 0.39786027127426865 y pvalor 0.002628949607816149, H0 no estan correlacio
nadas

Probablemente NO estan correlacionados
Test pearson para Georgia, estadistico 0.5446619100188149 y pvalor 6.828305061520844e-06, H0 no estan correlaci
onadas

Probablemente NO estan correlacionados
Test pearson para Germany, estadistico 0.5368738523807763 y pvalor 4.018864397349724e-06, H0 no estan correlaci
onadas

Probablemente NO estan correlacionados
Test pearson para Greece, estadistico 0.5641819868327069 y pvalor 3.6508282561000874e-07, H0 no estan correlaci
onadas

Probablemente NO estan correlacionados
Test pearson para Hungary, estadistico 0.5483915331470056 y pvalor 3.5203819445550243e-07, H0 no estan correlac
ionadas

Probablemente NO estan correlacionados
Test pearson para Ireland, estadistico 0.5218468697550063 y pvalor 6.903295748023087e-07, H0 no estan correlaci
onadas

Probablemente NO estan correlacionados
Test pearson para Italy, estadistico 0.5265028753436503 y pvalor 2.2717745537585754e-07, H0 no estan correlacio
nadas

Probablemente NO estan correlacionados
Test pearson para Latvia, estadistico 0.5329647460324982 y pvalor 6.384710847552133e-08, H0 no estan correlacio
nadas

Probablemente NO estan correlacionados
Test pearson para Lithuania, estadistico 0.5425506512299001 y pvalor 1.3515158090836495e-08, H0 no estan correl
acionadas

Probablemente NO estan correlacionados
Test pearson para Luxembourg, estadistico 0.5457401212982221 y pvalor 4.306543873845531e-09, H0 no estan correl
acionadas

Probablemente NO estan correlacionados
Test pearson para Moldova, estadistico 0.5496903637280637 y pvalor 1.2551812066563193e-09, H0 no estan correlac
ionadas

Probablemente NO estan correlacionados
Test pearson para Netherlands, estadistico 0.5332608624793453 y pvalor 2.0004400729517244e-09, H0 no estan corr
elacionadas

Probablemente NO estan correlacionados
Test pearson para North Macedonia, estadistico 0.45801141175391696 y pvalor 2.650131561907683e-07, H0 no estan
correlacionadas

Probablemente NO estan correlacionados
Test pearson para Poland, estadistico 0.4592878049300589 y pvalor 1.3172950771686062e-07, H0 no estan correlaci
onadas

Probablemente NO estan correlacionados
Test pearson para Portugal, estadistico 0.4471694188056265 y pvalor 1.7141099899520709e-07, H0 no estan correla
cionadas

Probablemente NO estan correlacionados
Test pearson para Romania, estadistico 0.4536104170110994 y pvalor 5.964375729631549e-08, H0 no estan correlaci
onadas

Probablemente NO estan correlacionados
Test pearson para Serbia, estadistico 0.43667747024043435 y pvalor 1.1913716221117234e-07, H0 no estan correlac
ionadas

Probablemente NO estan correlacionados
Test pearson para Slovakia, estadistico 0.43900653900985354 y pvalor 5.7677875536642705e-08, H0 no estan correl
acionadas

Probablemente NO estan correlacionados
Test pearson para Slovenia, estadistico 0.4387461726381103 y pvalor 3.39066709286736e-08, H0 no estan correlaci
onadas
```

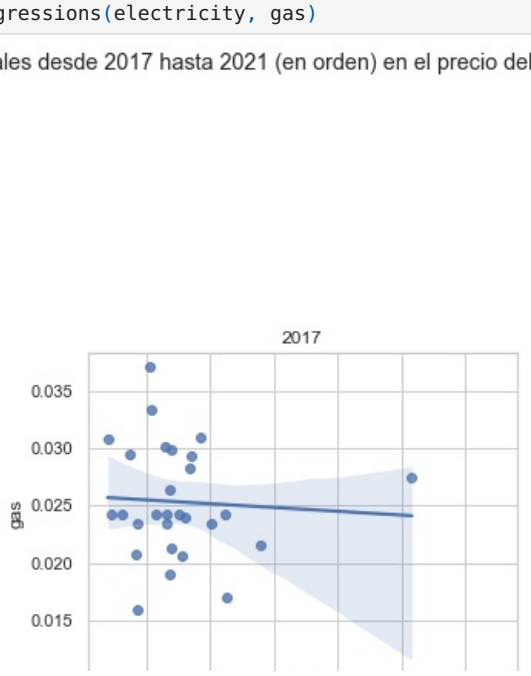
```
Probablemente NO estan correlacionados
Test pearson para Spain, estadistico 0.369728498662823 y pvalor 3.2173755292893473e-06, H0 no estan correlacion
adas

Probablemente NO estan correlacionados
Test pearson para Sweden, estadistico 0.2521460028362418 y pvalor 0.0015503878409114694, H0 no estan correlacio
nadas

Probablemente NO estan correlacionados
Test pearson para Turkey, estadistico 0.27413483078256173 y pvalor 0.00045179839914571816, H0 no estan correlac
ionadas

Probablemente NO estan correlacionados
Test pearson para Ukraine, estadistico 0.27310465067688705 y pvalor 0.00038649595749785427, H0 no estan correla
cionadas

Probablemente NO estan correlacionados
Test pearson para United Kingdom, estadistico 0.282540907419713 y pvalor 0.00018917062353675837, H0 no estan co
rrelacionadas

Probablemente NO estan correlacionados
```

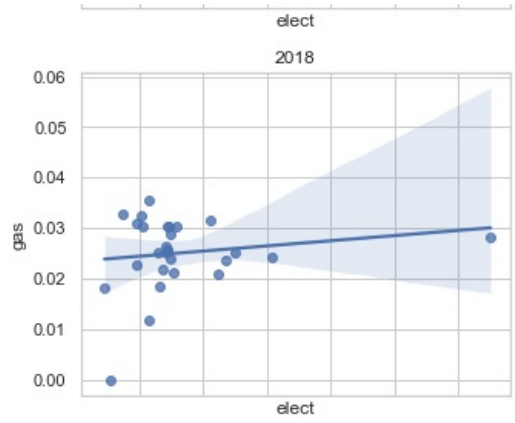
Tal y como se previó en el análisis visual rápido, se ha dado correlaciones en el gas y la electricidad para ciertos países, no obstante, seguimos teniendo heterocedasticidad en las muestras.

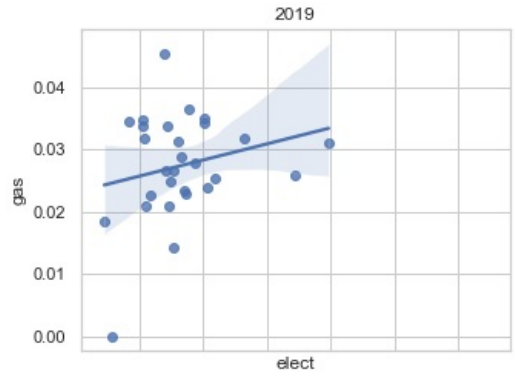
5. Regresiones

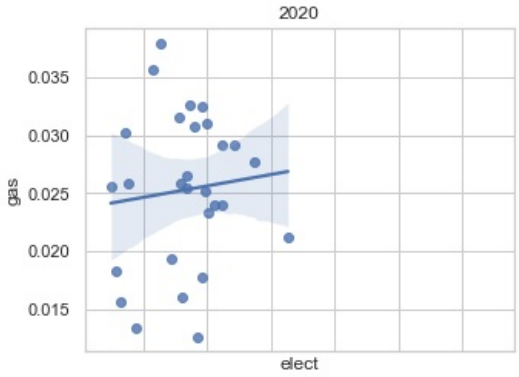
In [282…
```python
def plot_linear_regressions(electricity, gas):
    elec_2017 = pd.concat([electricity["2017"], gas["2017"]], axis = 1, keys = ["elect", "gas"])
    elec_2018 = pd.concat([electricity["2018"], gas["2018"]], axis = 1, keys = ["elect", "gas"])
    elec_2019 = pd.concat([electricity["2019"], gas["2019"]], axis = 1, keys = ["elect", "gas"])
    elec_2020 = pd.concat([electricity["2020"], gas["2020"]], axis = 1, keys = ["elect", "gas"])
    elec_2021 = pd.concat([electricity["2021"], gas["2021"]], axis = 1, keys = ["elect", "gas"])

    fig, axs = plt.subplots(6, 1, figsize=(5, 27), sharex=True)

    fig.suptitle("Regresiones lineales desde 2017 hasta 2021 (en orden) en el precio del gas y de la electricid
    sns.regplot(x = "elect", y = "gas", data = elec_2017, ax = axs[0])
    sns.regplot(x = "elect", y = "gas", data = elec_2018, ax = axs[1])
    sns.regplot(x = "elect", y = "gas", data = elec_2019, ax = axs[2])
    sns.regplot(x = "elect", y = "gas", data = elec_2020, ax = axs[3])
    sns.regplot(x = "elect", y = "gas", data = elec_2021, ax = axs[4])

    years = ["2017", "2018", "2019", "2020", "2021"]
    d1 = np.array([])
    d2 = np.array([])
    for d in years:
        d1 = np.append(d1, electricity[d].values)
        d2 = np.append(d2, gas[d].values)
    sns.regplot(x = d1, y = d2, data = elec_2021, ax = axs[5])

    axs[0].set_title("2017")
    axs[1].set_title("2018")
    axs[2].set_title("2019")
    axs[3].set_title("2020")
    axs[4].set_title("2021")
    axs[5].set_title("Global")
```

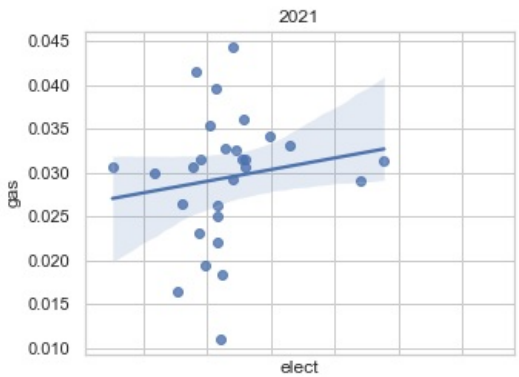
In [283…
```python
plot_linear_regressions(electricity, gas)
```

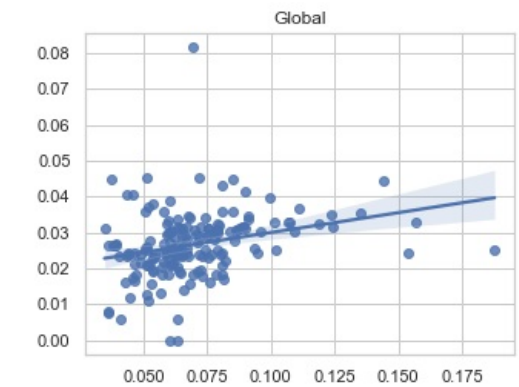












Tal y como vemos los regresores lineales, no generan un buen comportamiento, ni anual ni global, dentro de lo esperado por los problemas tanto generales como por país, de normalidad y heterocedasticidad. Realizar los regresores por país, carece de lógica por la poca cuantía de la muestra.

6. Conclusiones

6. Conclusiones

Ha quedado demostrado que la muestra a nivel de regresores, y por lo tanto uso para predicciones, no es recomendable, al menos con el tamaño y calidad actuales. Tenemos que tener en cuenta, que es tremendamente poder tener cifras exactas de todos los países de la UE historicamente, con un tamaño reseñable para conseguir estos objetivos, pues la digitalización de estos sectores, en algunos países es reciente. Así mismo, hemos podido comprobar que si existe correlación entre los precios de Gas y Electricidad en los países pertenecientes a la Zona Euro, pero no para la Comunidad de Países de la Eurozona,y además ambas muestras se comportan dentro de la normal.

In [ ]: