"Final Report"
Final Assignment CO2 Emissions

Salam Alsaadi

31-10-2022

Introduction

For this final assignment "CO2 Emissions", I diligently studied the requirements. It took me hours to orient myself in this subject. The website "Our World in Data" itself was immensely useful for me. So much scientific information and so many datasets and graphs to learn.

After all, I chose the following six datasets from this site:

1. owid-energy-data.csv
2. owid-energy-codebook.csv
3. co2-emissions-and-gdp.csv
4. kaya-identity-co2.csv
5. owid-co2-data.csv
6. levelized-cost-of-energy.csv

After data collection, I started analyzing dataframes and their columns to check the data quality (DQ) and whether they contain missing values in order to estimate needed efforts and steps for cleaning data and solve missing data problems, if any.

As a matter of fact, I have created DQ-functions to trace and improve DQ of 'my' datasets. These functions are "analyzeDataframes" and "find_missing_values".

The tactics I used here are:

A. looking at all the unique values
B. sorting and looking at the edges
C. casting to a type
D. looking at the frequency

Furthermore, my calculations are based on several assumptions/choices I made per question (of the totally 3 central questions of this assignment Q1-Q3).

Please see all my assumptions/choices and the reasoning behind them in this report as well as in my Python code.

Questions and answers

Q1) *What is the biggest predictor of a large CO2 output per capita of a country? [Biggest predictor of CO2 output]?*

Answer:

Here, I have implemented the following analysis:

'CO2 output per capita' versus [candidate Predictors of CO2 emissions].

The following two groups of candidates are examples of those predictors:

Candidates I: GDP per capita, diets, number of cars per capita, various energy source, mobility and other factors.

Candidates II: Population, GDP per capita, 'energy_per_gdp', 'co2_per_unit_energy', 'co2_per_gdp'.

For my analysis, I have chosen the second group (Candidates II).

The analysis columns in my selected group are categorized in this manner:

[main column] = 'annual_co2_emissions'
[candidate predictor columns] = ['population', 'gdp_per_capita', 'co2_per_gdp', 'energy_per_gdp', 'co2_per_unit_energy'].

In order to generate statistics for certain columns, I wrote for this purpose a special Python function named "generateStatistics(my dataframe, my analysis columns)".

After generating statistics for all above-mentioned columns (possible predictors versus main column), it was very obvious that the calculated correlation is –generally speaking- **relatively weak**; except the factor "gdp_per_capita" which scored [correlation coefficient= 0.984548] higher than the others.
P.S.> You will see later, that the formulas (2 methods) have scored considerably higher.

Here is an example of the weak correlation:

```
1) Correlation Coefficient using Pandas itself:
                      co2_per_unit_energy  annual_co2_emissions
co2_per_unit_energy              1.000000              0.051652
annual_co2_emissions             0.051652              1.000000

2) Correlation Coefficient using pingouin library:
            n         r          CI95%         p-val     BF10     power
pearson  9958  0.051652  [0.03, 0.07]            0.0  7470.51  0.999309
```

As it is shown, the correlation is very low [0.051652]. The p-value emphasizes this result too [value= 0.0] !

In fact, this undeniable conclusion has opened the door for other possibilities!

Therefore, I studied carefully two important scientific equations for calculating CO2 Emissions, where after I have applied them to generate combined correlation coefficients!

These two methods are:

Method I =
[CO2 emissions = Population * (GDP / Population) * (CO2 emissions per $)]

Method II =
[CO2 Emissions = Population * (GDP / Population) * (Energy / GDP) * (CO2 / Energy)]

If we translate this to concrete (involved) columns:

Method I = ['annual_co2_emissions' vs. { 'population' * 'gdp_per_capita' * 'co2_per_gdp' }]

Method II = ['annual_co2_emissions' vs. { 'population' * 'gdp_per_capita' * 'energy_per_gdp' * 'co2_per_unit_energy' }]

Finally, I created statistics using both formulas.

The results are definitely an improvement in the embodied correlation!
This leads us to the conclusion that applying these two formulas are a convincing proof of considering a real predictor for the CO2 output of a country (or for the whole planet).

Especially the first formula has scored extremely high in calculating the correlation between CO2 emissions and all involved factors.

Here is the outcome of statistical calculation using formula I:

```
1) Correlation Coefficient using Pandas itself:
                      co2_driver_method_1  annual_co2_emissions
co2_driver_method_1              1.000000              0.999905
annual_co2_emissions             0.999905              1.000000

2) Correlation Coefficient using pingouin library:
            n         r       CI95%  p-val BF10  power
pearson  8533  0.999905  [1.0, 1.0]    0.0  inf    1.0
```

Formula II on the other hand has scored a fraction less than formula I:

```
1) Correlation Coefficient using Pandas itself:
                      co2_driver_method_2  annual_co2_emissions
co2_driver_method_2              1.000000              0.999878
annual_co2_emissions             0.999878              1.000000

2) Correlation Coefficient using pingouin library:
            n         r       CI95%  p-val BF10  power
pearson  7109  0.999878  [1.0, 1.0]    0.0  inf    1.0
```

Conclusion

The biggest predictor of a large CO2 output per capita of a country is a combination of factors according to Method I (followed by Method II).

Predictor Method I = [Population * (GDP / Population) * (CO2 emissions per $)]

Q2) Which countries are making the biggest strides in decreasing CO2 output? [Biggest strides in decreasing CO2 output]?

In order to answer this question, I have used the relative CO2 output for each country.

I took into account that countries can vary in population every year. Their population can grow, but it can shrink too!

Besides, I have filtered the data in order to exclude data older than the year 1990. The reason for that is the poor Data Quality of the data before 1990.

In my solution, I have used two methods as follow:

Method I:
Calculate the biggest 10 strides in decreasing CO2 output internationally **[on country basis]**.

Overview result:

```
*********************************************************************************
[Biggest 10 strides in decreasing CO2 output internationally (on country basis)]:
(10, 4)
*********************************************************************************
                             country  year co2_per_capita  biggest_stride_decreasing_co2
0                            Curacao  1998          1.758                        -29.480
1          Sint Maarten (Dutch part)  1998          0.896                        -15.488
2  Bonaire Sint Eustatius and Saba  1998          0.862                        -14.559
3                              Aruba  2012         13.146                        -11.340
4                              Qatar  1998         57.512                        -11.212
5                             Kuwait  1991          7.343                        -10.700
6                            Bahrain  2001         20.182                         -8.096
7                          Singapore  2010         11.035                         -7.106
8               United Arab Emirates  2002          24.06                         -6.397
9                         Luxembourg  1995         22.423                         -6.258
*********************************************************************************
```

Method II:
Calculate the biggest 10 strides in decreasing CO2 output internationally **[on yearly basis]**.

Overview result:

```
********************************************************************************
[Biggest 10 strides in decreasing CO2 output internationally (on yearly basis)]:
(10, 4)
********************************************************************************
   year  country co2_per_capita  biggest_stride_decreasing_co2_year
0  1998  Curacao          1.758                             -29.480
1  2010  Curacao         26.054                             -12.766
2  2012    Aruba         13.146                             -11.340
3  1991   Kuwait          7.343                             -10.700
4  2007    Qatar         51.624                             -10.170
5  2017  Curacao         27.324                              -8.689
6  2001  Bahrain         20.182                              -8.096
7  2016  Curacao         36.013                              -7.309
8  2008    Qatar         44.726                              -6.898
9  2013   Kuwait         23.169                              -6.714
********************************************************************************
```

Moreover, aiming at an objective resu;ts comparison between old and relatively recent data, I have repeated the same steps to produce two additional tables for a recent period, namely from the year 2015 till the end.

Method I:
Calculate the biggest 10 strides in decreasing CO2 output internationally **[on country basis]** for the period starting from year 2015.

Overview result:

```
******************************************************************************
[Recent (from 2015), Biggest 10 strides in decreasing CO2 output internationally
(on country basis)]:
(10, 4)
******************************************************************************
                   country  year co2_per_capita  biggest_stride_decreasing_co2
0               Montserrat  2016          5.888                         -5.917
1                   Brunei  2015         16.761                         -5.020
2      Trinidad and Tobago  2016         29.075                         -3.853
3                     Niue  2015          4.552                         -2.288
4                    Libya  2015          9.065                         -2.156
5                  Iceland  2020          8.604                         -1.869
6            United States  2020         14.238                         -1.734
7                    Malta  2015          4.016                         -1.675
8   British Virgin Islands  2017            5.7                         -1.539
9                Hong Kong  2020          4.167                         -1.453
******************************************************************************
```

Method II:
Calculate the biggest 10 strides in decreasing CO2 output internationally **[on yearly basis]** for the period starting from year 2015.

Overview result:

```
******************************************************************************
[Recent (from 2015), Biggest 10 strides in decreasing CO2 output internationally
(on yearly basis)]:
(6, 4)
******************************************************************************
   year  country co2_per_capita  biggest_stride_decreasing_co2_year
0  2017  Curacao         27.324                              -8.689
1  2016  Curacao         36.013                              -7.309
2  2015   Brunei         16.761                              -5.020
3  2019  Estonia          9.339                              -4.225
4  2018  Curacao         23.526                              -3.798
5  2020    Qatar         37.019                              -3.600
******************************************************************************
```

Conclusion

The countries, which are making the biggest strides in decreasing CO2 output (units= tonnes of carbon dioxide per capita in a certain year), are a.o.

```
                         country  year co2_per_capita  biggest_stride_decreasing_co2
0                        Curacao  1998          1.758                        -29.480
1      Sint Maarten (Dutch part)  1998          0.896                        -15.488
2  Bonaire Sint Eustatius and Saba  1998          0.862                        -14.559
```

Q3) **Which non-fossil fuel energy technology will have the best price in the future? Best future price for non-fossil fuel energy?**

I have used the dataset "Levelized-cost-of-energy" from the website "Our World in Data". Then, I filtered the dataset on country name 'World' as a representative for all countries together!

Applying a linear regression, I have made a statistical model over the various non-fossil fuel options based on current data.

Thereafter, by using my model, I have predicted future prices for those non-fossil energy technologies.

Finally, I also used scatterplot for displaying the graphs between time and prices.
This goes for current prices as well as for predicted future energy prices.

P.S. Temporarily, I do not execute the Py-code for running the scatterplot for every item.
Instead, I let the code to run with one item as a demo!

The result of my analysis is:

```
*************************************************************************
[Top 3 Best future price for non-fossil fuel energy]:
*************************************************************************
  country  year pred_non_fossil_energy_tech  pred_future_price_in_2035
0   World  2035            pred_solar_price                  -0.457320
1   World  2035              pred_csp_price                  -0.094208
2   World  2035     pred_onshore_wind_price                  -0.007526
*************************************************************************
```

<u>Conclusion</u>

The top 3 non-fossil fuel energy technologies, which will have the best price in the future, are as follow:

1) Solar energy as a whole.
2) CSP (Concentrated Solar Power) plants.
3) Onshore wind energy.