Individual Delivery

23/12/2020 to 12/01/2021

—

Antonio Jesús Leal Martín
Data Science
The Bridge

General Vision

It is a fact that the Earth temperature has increased during the last century. In particular, and according to to the Goddard Institute for Space Studies (GISS), the average global temperature has increased 0,8ºC during the last century. This temperature change comes with some impacts on the planet, like the melting of the glaciers or the rise of sea levels.

This is mostly caused by the increasing emission of gases that absorb the heat from Earth surface radiation. Among all those gases, we will focus on the study of the CO2 emissions, as it is the gas whose emissions increased the most since the Industrial Revolution, due to the use of fossil fuels.

Goals

My goal for this delivery is **A+**. A requirement check could be found below:

General

1. Give an answer to a hypothesis. ✓
2. Make a presentation. ✓
3. Divide the tasks. ✓
4. Use Trello. ✓
5. Sent before 12/01/2021 at 23:59. ✓
6. Use the proposed folder structure. ✓

Option C

1. Document al steps. Write code using good practices. ✓
2. Use Trello. ✓
3. Collect the data. ✓
4. Determine if data is cleaned. If not, clean it. ✓
5. Show tendencies of each column in datasets. ✓

   A lineplot of each column of both dataframes (CO2 emissions and temperature) has been made and saved to "..\reports\plots\lineplots"

6. Create a pie chart to show the time needed for each step of the project. ✓
7. Answer questions: ✓

   a) Was it possible to demonstrate your hypothesis?

   The hypothesis has been refuted. With this data, there is no clear relationship between the local CO2 emission and the local increase of temperature.

   b) What can you conclude about your data study?

   Despite the hypothesis cannot be demonstrated some other facts can be extract from this study, like every country has an increasing tendency on the average yearly temperature, or that the countries that emit more CO2 are those that

posess the greatest oil wells and refineries.

c) What would you change if you need to do another EDA project?

I would change the topic, as I selected this one in a hurry (because the one that I chose first had not enough data). I has been very hard for me to find the motivation to do this analysis. On the other hand, I would use other visualization tools to represent the data.

d) What do you learn from this project?

I have learned to search data, to organize and clean it, and to get it prepared for its posterior analysis. I have also learned to use some python modules like sklearn or basemap.

Option B

1. Show histograms from each column. How are the ranges painted? ✓

   Histograms of each column of the dataframes (CO2 emissions and temperature) has been made and saved to "..\reports\plots\histograms".

   To show how the ranges were painted, at the x axis label of each plot it is shown the bin width.

2. Which are the columns with the highest correlation? Draw correlation matrix. ✓

   Since the datasets have 123 columns each and making that big correlation matrix did not give any interesting information about the hypothesis, I made the correlation between the CO2 emission and Temperature for each country. Results of these correlation can be found at "..\reports\plot".

3. Use matplotlib to show the graphs. ✓

Option A

1. Save each plot in local files. ✓
2. Use distribute modules. Main Jupyter Noteebok only contains calls. ✓
3. Apart from matplotlib, use seaborn to show the graphs. ✓
4. Answer the questions: ✓

   a) Are there outliers or some rare data?

   No. I found some strange peaks on the CO2 emission values that at first made me suspect that there were some outlier values, but those peaks correspond to countries that have big oil reservers and refineries in the early 80's, when CO2 emission were not as regulated as it is today. They were correct values.

   b) What are the columns with more repeated values?

   Since the columns of both datasets show meassured data with decimals for CO2 emission and Temperature, the repetition of values is mere coincidence, in fact, no repeated value where found through the datasets columns.

Option A+

1. Create a pull request for the entire project. ✓
2. Are there more urls from where to collect your data? ✓

   Since the data I have made this study with reaches only 2013, I made a research to find a more updated dataset.
   Unfortunately, apart from the data of temperatures for some countries, I was not be able find any complete data that allow me to expand the main datasets.
3. Use classes to practice OOP. ✓

Specifications

Software

The minimum software requirements to run the attached code are the following:

- Python 3
- VSCode with Python and Jupyter extensions installed.
- Python modules: pandas, numpy, basemap, sklearn, seaborn, matplotlib.

Hardware

The software has been ran using the following computer specs:

- **CPU:** Intel(R) Core(TM) i7-6700 (at least Inte(R) Core(TM) i5 recommended)
- **RAM:** 16Gb (more than 8Gb recommended)
- **GPU:** AMD Radeon R9 390
- **OS:** Windows 10

Requirements

The required data to run this program and obtain the corresponding results is included in the folder "..\resources". In case you want to donwload it from the source, you can get it from:

Temperature:

https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data

CO2 emission:

https://data.worldbank.org/indicator/EN.ATM.CO2E.PC

A cleaned version of de datasets has been also included with the project at the folder. You cand find it at the folder "..\resources\clean_data"

Steps

I. Research the context

The first task of this project was the research for the topic and data to be analyzed. After failing first (I chose a project with not enough amount of data), I consulted Kaggle dataset collection in order to find a topic that draw my attention.

I have always been concerned about climate change and its consecuences, so when I found this temperature dataset on Kaggle I decided to use it for this project. At first, my hypotesis were to relate temperature changes with the human developement index (HDI) given for each country by the United Nations Organization. But, since that index has not direct relationship with the atmospherical phenomena, I decided to move into contamination, more particularly, to CO2 emissions, and check if the local temperature change can be correlated with the local CO2 emissions for each country.

II. Get Data

As I explained in the "Requirements" section above. I directly got the data in csv format from the following sources:

Temperature:

https://www.kaggle.com/berkeleyearth/climate-change-earth-surface-temperature-data

CO2 emission:

https://data.worldbank.org/indicator/EN.ATM.CO2E.PC

Also, for the making of some map plots, I got an small csv document with the coordinates (latitude and longitude) of all the contries from:

https://developers.google.com/public-data/docs/canonical/countries_csv

III. Data Wrangling

I changed the format in wich the information was presented, to have a more efficient way to work with it. For both datasets, I wrangled them to get the folowing structure:

- **Columns:** Names of the conuntries ordered alphabetically.
- **Index:** Dates in ascending order (datetime format)
- **Content:** The corresponding CO2 or temperature data.

In this way, having the two dataset sharing a common structure, it was way easier to work with them and making the corresponding analysis

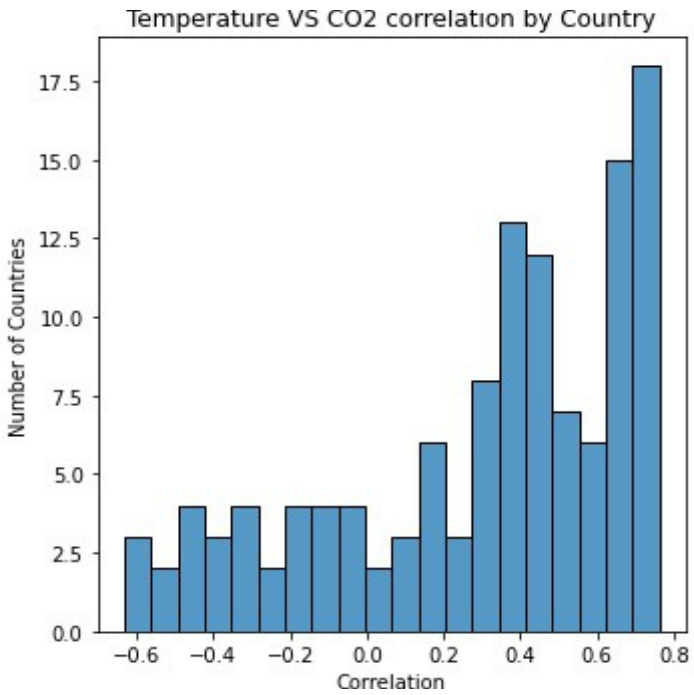
IV. Data Mining / Clean Data

I fond that both dataset contained incomplete data (NaN) values, for some countries and dates. First of all, from both datasets I removed all the countries with any NaN value in the data. After that I compared both datasets to check the common countries, and removed the data from the rest of the countries as it was not possible to make the relationship between CO2 emissions and temperatures.

Finaly, I did only kept the dates from 1960 to 2016 for the CO2 data, and from 1910 to 2013 for the temperature dataset.

V. Analysis

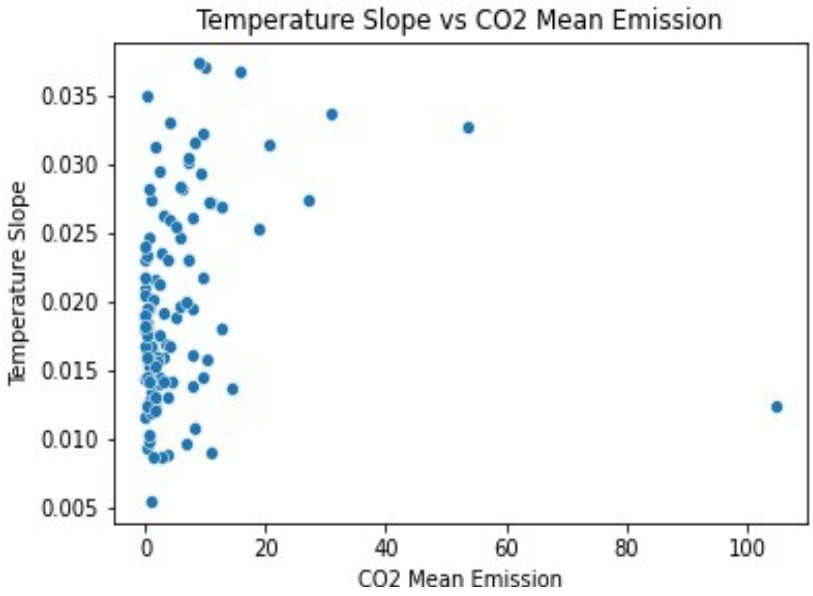
In order to find any relationship between temperatures and CO2 emission, I started by calculating the correlation between CO2 data column and temperature data column for each country, getting the following



Only 51 countries were found in the 'relevant' correlation range (c > 0.5 and c < -0.45). That fact shows that the correlation between both dataset is not strong enough to consider it a rule.

I also tried to establish a relationship between the temperature increase rate and the CO2 average emission for each country. To do so, I calculated (using mklearn) the linear regression for the temperatures of every country from 1960 to 2013 (I used this range to couple with the CO2 data), and I got the slope of those refressions, getting some info about the temperature increase rate.

When I compared those slopes with the average CO2 emissions for that time range, I found the results below:



As it can be seen, this plot shows that there is no direct relationship between a higher average CO2 emission and a steeper slope of the temperature tendency.

VI. Conclusion

As seen in the previous section, we can conclude that the hyphothesis that we tried to demonstrate was not true. It has been scientifically demonstrated that CO2 (and other greenhouse effect gases) emission to the atmosphere has consecuences on the global warming, but it does not seem to affect locally. We can find regions with high CO2 emissions and low temperature increase tendency and vice-versa.

VII. Project Scope

There are some aspects not analyzed in this project that could be interesting to analize in the future:

- Analize temperature and CO2 emissions taking into account the country and its neighbours
- Consider the location of each country and search for correlations between temperature, emissions and coordinates (longitude and latitude)
- Try to predict future temperatures based in known data.