Patterns and Predictions of CO2 Emissions: A Statistical Analysis of Historical Data from Portugal and the World (1900-2021)

José Teixeira  
*Engenhária Informática*  
*Instituto Superior de Engenharia do Porto*Porto, Portugal  
1200941

João Morais  
*Engenhária Informática*  
*Instituto Superior de Engenharia do Porto*Porto, Portugal  
nmr

*Abstract* - This study, utilizing the Python programming language, conducts analysis of CO2 emissions in Portugal and other selected regions, covering the period from 1900 to 2021. Using data analysis techniques, including exploratory analysis, statistical inference, correlation, and regression modeling, it compares per capita emissions between Portugal and Spain, and the impact of various emission sources such as cement, coal, burning, gas, and oil. Additionally, the correlation between CO2 emissions from the selected regions is examined, and a linear regression model is developed to predict emissions in Europe, based on emissions from key countries.

# Introduction

Carbon dioxide (CO2) emissions are among the major contributors to climate change, significantly impacting the environment and society. Understanding historical trends, emission sources, and relationships between different regions becomes crucial for formulating effective mitigation strategies. This study focuses on analyzing CO2 emissions in Portugal and other selected regions from 1900 to 2021, using a multidisciplinary approach that combines exploratory data analysis, statistical inference, correlation, and regression. The aim is to provide a comprehensive view of CO2 emission dynamics, identify patterns, and predict future trends, thereby contributing to the development of more informed and effective environmental policies.

To achieve the proposed objectives, the study employed several advanced statistical techniques, detailed as follows:

**Exploratory Data Analysis (EDA)**: EDA was used to gain an initial understanding of the data, including the distribution, trends, and patterns of CO2 emissions. This phase included data visualization through line charts and histograms to represent total CO2 emissions, as well as the analysis of emissions by specific source and per capita comparisons between Portugal and Spain.

**Statistical Inference**: Statistical inference was used to test hypotheses about the gross domestic product (GDP) averages between Portugal and Hungary and to examine the differences in total CO2 emissions between regions. T-tests for independent samples and analysis of variance (ANOVA) were applied, followed by post-hoc analyses when necessary, to evaluate statistically significant differences between groups.

**Correlation**: Correlation analysis was used to explore the relationships between CO2 emissions from different regions, identifying possible interdependencies. The analysis was performed by calculating correlation coefficients, which were visualized through a correlation matrix.

**Linear Regression**: A linear regression model was developed to investigate the relationship between CO2 emissions in specific countries and total emissions in Europe, using selected independent variables. This model allowed assessing the influence of each country's emissions on European emissions, checking the model's adequacy through residual analysis and multicollinearity (VIF), and making predictions about future emissions.

This study combines these statistical techniques to provide a detailed analysis of CO2 emissions, highlighting the importance of an integrated approach in understanding global environmental issues. By applying these methodologies, the work not only identifies patterns and trends in CO2 emissions but also offers insights for evidence-based mitigation policy formulation.

# Data analysis and exploration

## CO2 Emissions Visualization: Portugal’s total in the period 1900-2021

Utilizing the “pandas” [1] library, we conducted a reading of the provided data [2]. Furthermore, in conjunction with “matplotlib” [3] library, in order to create the graphical representation *Figure 1*, trough wich a time series analysis was performed.

This technique facilitates the visualization of the evolution of Portugal's total CO2 emissions over the period from 1900 to 2021. The time series is depicted graphically, with the x-axis denoting the year and the y-axis quantifying the total CO2 emissions in MtCO2 (Megatons of CO2).

Employing the “idmax” [1] function from the library “pandas” [1], we ascertained the year with highest emissions in Portugal, which was identified as 2005 with a value of 69.71 Megatons as depicted in *Figure 2*.

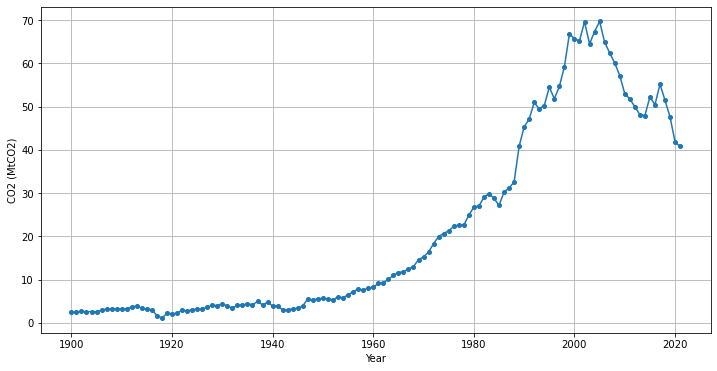


Figure 1. Total CO2 Emissions in Portugal (1900-2021) Megatons of CO2 (Mt) per Year

|  |  |
| --- | --- |
| Year | MtCO2 |
| 2005 | 69.718 |

Figure 2. Table of Data for the dataframe: informacao\_ano\_max\_emissoes

**Early 20th Century (1900-1950)**: There was a relatively stable and low level of emissions, which suggests minimal industrial activity or the use of less carbon-intensive technologies and fuels during this period.

**Mid 20th Century (1950-1970)**: A gradual increase began around 1950, indicative of increased industrialization, economic growth, or changes in energy production and consumption patterns.

**Late 20th Century (1970-2000)**: A period of rapid and consistent growth in emissions is observed, peaking towards the end of the century. This could be associated with substantial industrial expansion, increased reliance on fossil fuels, and higher energy consumption due to economic development and a growing population.

**Early 21st Century (2000-2010)**: The ascending trend continues, reaching an all-time high just after the year 2000. It's possible that this reflects a culmination of previous growth trends before any significant environmental policies or technology shifts took effect.

**Recent Decade (2010-2021)**: Post-2010 shows a volatile but generally declining trend, with noticeable dips and recoveries. This could be influenced by a variety of factors, including economic crises, policy measures to reduce carbon emissions, the adoption of renewable energy sources, or improvements in energy efficiency.

## Sector CO2 Emissions: Portugal in the period 1900-2021

Once again, utilizing the “pandas” [1] library, we conducted a reading of the provided data, alongside “matplotlib” [3] library, for the creation of the graphical representation Figure 3.

In this analysis, we observe the varying quantities of MtCO2 (Megatons of CO2) emitted yearly in Portugal within the specified temporal frame by different sources: cement, coal, burning, gas, and oil.

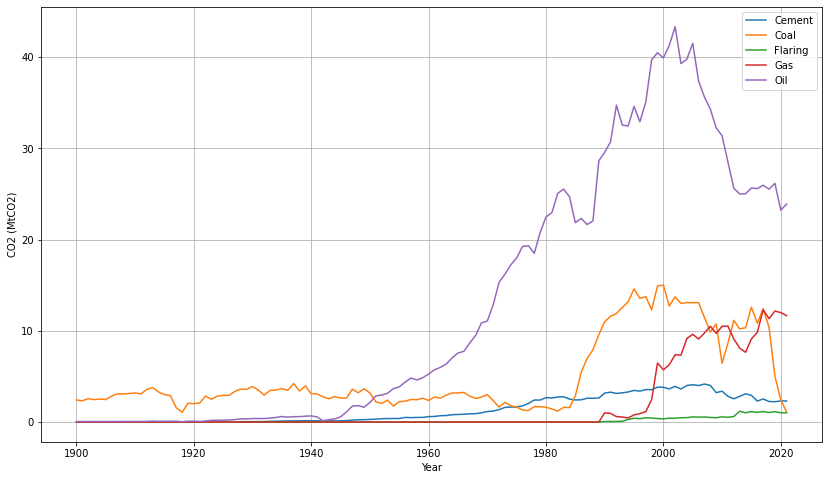


Figure 3. CO2 Emissions from Different Sources in Portugal (1900-2021) Megatons of CO2 (Mt) per Year

**Cement - Blue Line**: The emissions from cement production show a modest increase starting from around the 1950s. A steep increase is noticeable from the late 20th century, indicating a booming construction industry. The peak in emissions occurs around the early 2000s, followed by a slight decrease, possibly due to improved technologies or environmental regulations.

**Coal - Orange Line**: Coal emissions remain low until the 1950s, suggesting limited use in earlier years. There is a consistent rise from the 1950s through to the late 20th century, which then levels off and eventually decreases, likely reflecting a shift away from coal as an energy source.

**Fires - Green Line**: Emissions from fires are relatively stable and low throughout the century, with some fluctuations that could be attributed to variable rates of deforestation, land clearing, or wildfire incidents.

**Gas - Red Line**: Gas emissions remain low until the 1970s, after which they increase steadily. The trend in gas emissions climbs particularly sharply in the 1990s and continues to rise, suggesting a growing reliance on natural gas.

**Oil - Purple Line**: Oil emissions rise significantly post-1950, indicating its growing importance as an energy source during Portugal’s industrialization. There is a sharp increase peaking around 2000, followed by a pronounced decrease, which might indicate the impact of alternative energy sources or energy efficiency measures taking effect.

## CO2 Emissions : A Comparative Analysis of Portugal and Spain per capit 1900-2021

Following the same procedure regarding the reading of the provided data [2] and creation of a graphical representation [3], we mapped the emissions of CO2 per capita for Portugal and Spain over a specified period (1900-2021).

Initially, we filtered the dataset to isolate the records for Spain, enabling a focused analysis on the country's data. Subsequently, for both Portugal and Spain, we calculated the CO2 emissions per capita, converting the values into metric tons for standardization and comparability.

Equation 1. CO2 Emissions Per Capita Formula

This calculation was achieved by dividing the total CO2 emissions by the population count of each country, then multiplying by a factor of one million to convert the results into metric tons.

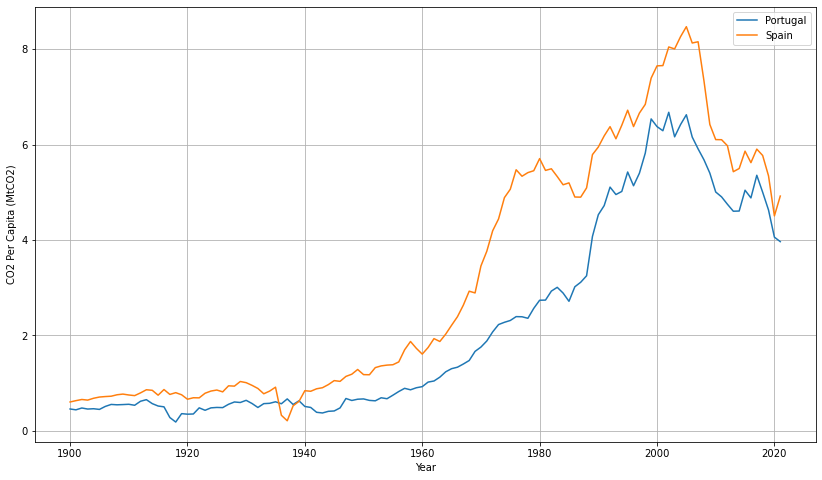


Figure 4. CO2 Emissions per Capita Portugal vs Spain (1900-2021) Megatons of CO2 (Mt) per Year

In comparison, Spain's per capita emissions consistently exceed those of Portugal for much of the period from 1900 to 2021. Notably, in the 1930s, there is an exception where Spain's per capita emissions fall below Portugal's, likely influenced by the economic impacts of the Spanish Civil War [4], leading to reduced industrial activity and energy consumption. After this period, from the 1950s onwards, Spain's emissions increase more pronouncedly than Portugal's, suggesting a more rapid pace of industrialization or energy consumption growth. The peaks in emissions per capita are higher in Spain and occur slightly later than in Portugal. Post-peak, the decline in Spain's emissions is sharper, which may reflect more aggressive environmental policies or a significant restructuring in energy use and economic activities. Overall, while Spain has historically had a higher carbon footprint per person, it has also experienced a more notable decrease in emissions in recent years.

## Coal-Driven CO2 Emissions: A Comparative Analysis of the US, China, India, EU, and Russia, 2000-2021

Following a similar methodology applied in previous sections, we conducted a comparative analysis of coal-derived CO2 emissions for five major regions: the United States, China, India, the European Union (27), and Russia over the period from 2000 to 2021.

The data [2] were carefully filtered to extract CO2 emissions specifically generated from coal combustion within the targeted regions and time frame. This subset of data provided a precise foundation for our comparison. We visualized the emissions trends using a line graph that effectively displays the trajectory of each region's coal-related CO2 emissions over the two decades.

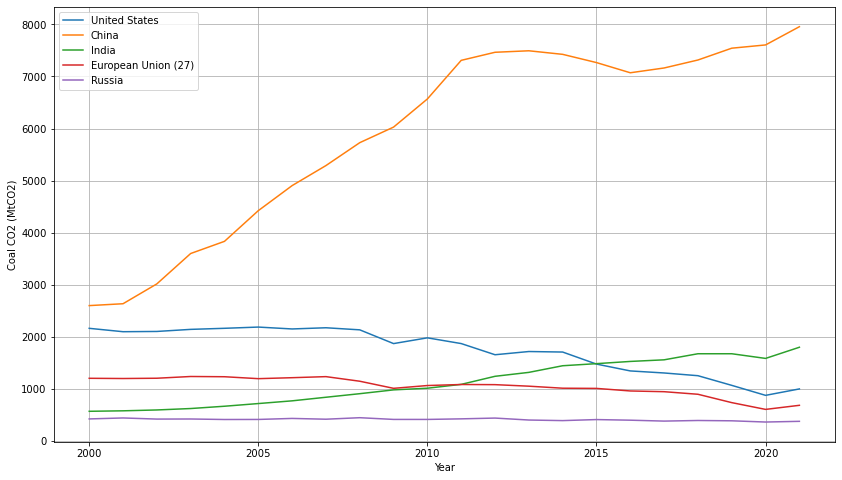


Figure 5. CO2 Emissions from United States, China, India, Europe Union, Russia (2000-2021) Megatons of CO2 (MtCO2) per Year

**United States - (Blue Line)**: Indicates the CO2 emissions from coal combustion in the United States from 2000 to 2021.

**China - (Orange Line)**: Represents the steep increase in China's coal-driven CO2 emissions during the same period.

**India - (Green Line):** Shows a steady rise in India's CO2 emissions from coal, reflecting ongoing industrial growth.

**European Union (27) - (Red Line)**: Depicts the emissions within the EU, with variations indicating changes in energy policy and consumption.

**Russia - (Purple Line):** Illustrates the fluctuations in Russia's coal-related emissions, suggesting shifts in industrial output and energy strategies.

The graph *Figure 5* reveals significant disparities in emissions trends across the regions. China's upward trajectory is the most pronounced, reflecting its rapid industrial growth and extensive reliance on coal as an energy source. Meanwhile, emissions from the United States and the European Union display a different pattern, with a notable plateauing and, in some cases, a reduction, which may be attributed to a shift towards cleaner energy sources and more stringent environmental regulations. India shows a steady increase, whereas Russia's emissions have fluctuated, likely due to varying industrial activities and energy policies.

Upon examination, it is clear that China is the predominant emitter of CO2 from coal combustion, with its emissions surpassing the combined total of the other analyzed regions, a testament to its significant industrial output and reliance on coal energy.

## Sector CO2 Emissions: A Comparative Analysis of the US, China, India, EU, and Russia, 2000-2021

This section extends our methodology to a broader spectrum of CO2 emissions, we conducted a analysis of different sectors derived CO2 emissions for five major regions: the United States, China, India, the European Union (27), and Russia since 2000 to 2021 as a average.

The data [2] were carefully filtered by country and time frame. During the calculations the output’s were rounded by 3 decimals.

Given a dataset [2] with records for each country, where each record contains the CO2 emissions from cement, coal, flaring, gas, methane, nitrous oxide, and oil for a specific year, the mean for each emission source is calculated as follows [5]:

Equation 2. Arithmetic mean

the mean (average) emissions for a given source for a country.

represents the emission value for a given source in the i-th year.

is the total number of records for the country within the specified period (from 2000 to 2021).

Equation 1 is performed for each of the specified emission sources for each country in the list ('United States', 'China', 'India', 'European Union (27)', 'Russia').

| Region | CO2 Emissions (MtC02) | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Coment | Coal | Fires | Gas | Methane | Nitrous Oxide | Oil |
| China | 599.141 | 5920.797 | 1.722 | 287.021 | 1015.726 | 476.53 | 1116.257 |
| EU(27) | 81.488 | 1049.236 | 21.132 | 774.871 | 407.444 | 238.482 | 1374.161 |
| India | 91.512 | 1123.795 | 2.661 | 92.464 | 617.36 | 228.242 | 469.662 |
| Russia | 21.837 | 413.504 | 43.061 | 766.698 | 599.007 | 58.484 | 353.289 |
| United States | 40.055 | 1750.037 | 52.728 | 1364.198 | 639.154 | 259.03 | 2379.692 |

Figure 6. CO2 Emissions by Sector from United States, China, India, Europe Union, Russia (2000-2021) Megatons of CO2 (MtC02) 2) yearly average.

**China**: Dominates in coal CO2 emissions, with a staggering average that highlights its heavy dependency on coal for energy. Additionally, significant emissions from cement production and methane indicate broad environmental impacts.

**European Union (27)**: Exhibits a balanced emission profile with notably high averages in gas and oil emissions. This reflects the EU's diverse energy mix and its efforts to manage coal dependence.

**India**: Shows a consistent rise in coal emissions, underlining its ongoing industrial expansion. Methane emissions are also significant, pointing to considerable agricultural and energy-related methane sources.

**Russia**: Characterized by high gas emissions, suggesting a major reliance on natural gas for energy. Flaring emissions are notably high, indicative of oil and gas extraction practices.

**United States**: Presents a diversified emission profile with the highest oil emissions, alongside substantial coal and gas emissions. This mirrors the U.S.'s broad energy consumption patterns and its challenges in transitioning to cleaner energy sources.

China's role as the predominant emitter across multiple categories underscores the need for targeted policy interventions and technological advancements in clean energy. Meanwhile, the varied emission profiles of the other regions highlight the importance of a multifaceted approach to addressing climate change, incorporating renewable energy and energy efficiency.

# Statistical inference

## Random Sample Analysis of GDP Performance: A Comparative Study of Portugal and Hungary Using Selected Years (1900-2021)

Following the same procedure regarding the reading of the provided data [2] we filtered the dataset to isolate the records for Portugal and Hungary. Subsequently, a random sample of years was generated within the period of study.

Specifically, we set a random seed using numpy “random.seed()” function [6] to ensure that our results are consistent and can be replicated. We then created a series of years ranging from 1900 to 2021 using pandas [1] “series()” function. From this series, we selected a random sample of 30 years without replacement, ensuring that each year was unique and representative of the entire period. This sample was then sorted in ascending order, facilitating a chronological analysis of the GDP data for Portugal and Hungary across the selected years.

To statistically determine whether Portugal's average GDP exceeded that of Hungary over the sampled years, a two-sample t-test was utilized under the assumption of unequal variances. For this we utilized the library scipy [7] and it’s stats.ttest\_ind function which applies the following formula:

Equation 3. two-sample t-test formula

is the t-statistic

and are the sample means for Portugal and Hungary, respectively,

and denote the sample variances

and represent the sample sizes.

|  |  |
| --- | --- |
| t | p |
| 0.184 | 0.427 |

Table 1. t and p values

The t-value of 0.184 and p-value of 0.427 from our study imply a **minimal statistical difference** between Portugal and Hungary's average GDPs across the selected sample years. This low t-value suggests the difference in GDP averages is not significant, backed by the p-value which indicates **insufficient evidence to suggest a real disparity in economic performance**.

## Bilateral GDP Analysis Using Distinct Random Year Samples: A Comparative Study of Portugal and Hungary (1900-2021)

Adopting a similar approach to our first analysis and referencing the methodology used there, we proceeded with an alternative statistical examination to further explore the economic performance of Portugal and Hungary through their Gross Domestic Product (GDP) over different sampled years.

Leveraging the numpy library's random.seed() function [6], we established a new seed value to ensure the reproducibility of our sampling method. Using the pandas [1] Series() function [1], we generated a new sequence of years spanning from 1900 to 2021. From this sequence, in contrast to our previous method, we extracted two non-overlapping random samples of 12 years each for Portugal and Hungary without replacement, using distinct seed values for each to guarantee unique and unbiased representations of the study period.

To statistically assess whether the average GDP of Portugal was superior to that of Hungary based on these independently drawn samples, we employed a two-sample t-test assuming unequal variances, utilizing the scipy.stats.ttest\_ind function from the scipy library [7]. This test applies the Equation 3.

|  |  |
| --- | --- |
| t | p |
| 0.244 | 0.405 |

Table 2. t and p values

These results suggest a similarly minimal statistical difference between the average GDPs of Portugal and Hungary as observed in our first question's outcome, where the t-value was 0.244 and the p-value was 0.405. The consistency in findings across two different sets of sampled years **reinforces the conclusion that there is insufficient evidence to claim a significant disparity** in economic performance between the two countries over the sampled periods.

## Regional CO2 Emissions Analysis Through Random Year Sampling: A Comparative Study Across the United States, Russia, China, India, and the European Union (1900-2021)

We extended the scope of our research to investigate the significance of differences in CO2 emissions among major regions, namely the United States, Russia, China, India, and the European Union (27). We conducted this analysis by employing the years previously selected in the analysis before, ensuring a consistent temporal framework for comparison.

Initially, we prepared our dataset by filtering the CO2 emissions data to include only the years within “sampleyears2” and the aforementioned regions. The data was then aggregated to calculate the total emissions for each region within the sampled years. To accommodate this process, we utilized the pandas library in Python to facilitate data manipulation and grouping [1]. Given the absence of direct records for the European Union (27), we acknowledged the need for data adjustment, which could involve the aggregation of individual member states' data, provided that such records were available.

With the data thus curated, we implemented an **ANOVA** (Analysis of Variance) test to ascertain the presence of significant differences in emissions among the regions. For this, we made use of the f\_oneway function from the scipy library's stats module [7].

|  |  |
| --- | --- |
| f | p |
| 5.288 | 0.00335 |

Table 3. f and p values

The computation yielded an F-statistic of 5.288 and a highly significant p-value of 0.00335, indicating that variations in CO2 emissions across the regions were unlikely to be attributed to random variation.

Given the significant result from the ANOVA test, it was **pertinent to conduct a post-hoc analysis** to identify which specific pairs of regions exhibited significant disparities in emissions. To this end, we applied the Tukey Honest Significant Difference (HSD) test using the function from the statsmodels library [8] . This test is particularly suited for all possible pairwise comparisons while controlling for Type I error across multiple tests.

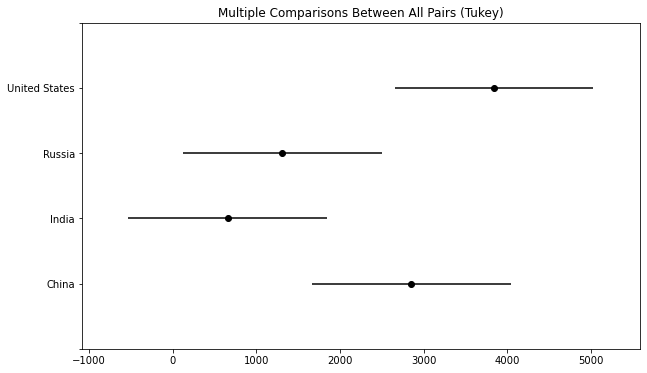


Figure 7. Tukey Honest Significant Difference (HSD) Test on CO2 Emissions

In conclusion, our findings substantiate the presence of significant disparities in total CO2 emissions among the selected regions during the years in sampleyears2. The ANOVA test's p-value strongly **rejects the null hypothesis** of equal means, while the Tukey HSD test offers a nuanced perspective on pairwise regional differences. The results from this analysis could serve as a scientific underpinning for region-specific policy-making and intervention strategies aimed at mitigating CO2 emissions.

# Correlation and Regression

## CO2 Emissions from Coal: A Correlation Table Analysis across Continents (2000-2021)

Continuing our investigation into the environmental impacts by region, we conducted a correlational study to analyze CO2 emissions specifically from coal between the years 2000 and 2021. We directed our focus on the interconnectedness of emissions between the regions of Africa, Asia, South America, North America, Europe, and Oceania.

In the preparatory phase of our data processing, we filtered our comprehensive CO2 emissions dataset to reflect only the relevant twenty-one-year period. We further refined our data to include emissions solely from coal. Utilizing pandas library [1], we grouped the data by region and year, summarizing the total coal-based CO2 emissions to reflect annual regional outputs.

After careful curation, we proceeded to transform our dataset into a format conducive to correlation analysis. The transformation process pivoted the years into rows and the specified regions into columns, enabling a straightforward computation of the correlation coefficients between the regional emissions. This resulted in a structured matrix that articulated the degree to which the emissions in one region could be statistically associated with those in another.

Leveraging seaborn and matplotlib libraries [9] [3], we produced a **heatmap** visualization to represent this correlation matrix.

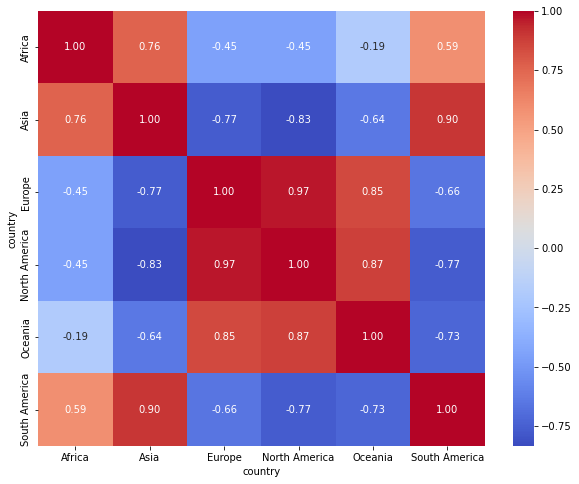


Figure 8. Heatmap

The resulting heatmap Figure 8 illuminated the presence of both strong and weak correlations between the region’s coal CO2 emissions over the given time frame. Certain regions demonstrated a higher degree of correlation, suggesting a potential linkage in their emission trends, while others showed little to no statistical relationship, underscoring the diversity of regional emission profiles.

A particularly strong positive correlation of **0.97** between **North America** and **Europe** indicates that as coal emissions rise or fall in one region, they are likely to do so in a similar fashion in the other. This may reflect shared economic activities or energy policies that closely align these regions' emissions profiles.

Conversely, **Asia** and **Oceania** exhibited a notably weaker correlation, suggested by a coefficient of **-0.64**. This implies that an increase in coal emissions in **Asia** does not necessarily correlate with an increase in Oceania, and may even inversely correlate, potentially due to differing energy consumption patterns or stages of industrial development.

**Africa's** correlation with other regions presented a mixed scenario; for instance, its correlation with **South America** stood at **0.59**, suggesting a moderate positive relationship, perhaps indicative of similar development strategies or energy dependencies. Yet, its correlation with **North America** was -**0.45**, pointing towards divergent trends or perhaps disparate energy sector dynamics between the two regions.

## Exploring Coal-Driven CO2 Emissions in Europe: A Multivariable Linear Regression Model (2000-2021)

Expanding upon our earlier analyses, we delved into the intricacies of coal CO2 emissions across various European countries during the even years of the 21st century. We focused on Germany, Russia, France, and Portugal as independent variables to model their influence on the dependent variable—Europe's aggregate emissions.

After filtering the CO2 data to include only the even years from 2000 to 2021, we structured our dataset around these specific variables. We ensured our analysis was grounded on solid statistical footing by employing a linear regression model using statsmodels library [8], where Europe's emissions served as the dependent variable, informed by the emissions data from the selected countries.

The following data represents the summary of such model:

|  |  |
| --- | --- |
| Metric | **Value** |
| Dep. Variable | Europe |
| R-squared | 0.984 |
| Adj. R-squared | 0.974 |
| F-statistic | 93.23 |
| Prob (F-statistic) | 1.57e-05 |
| Date | 03 Apr 2024 |
| Time | 23:07:45 |
| Log-Likelihood | -53.778 |
| No. Observations | 11 |
| AIC | 117.6 |
| BIC | 119.5 |
| Df Residuals | 6 |
| Df Model | 4 |
| Covariance Type | nonrobust |

Table 4. Model Summary

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Term** | **Coefficient** | **Standard Error** | **t-Statistic** | **P-value** | **95% Confidence Interval** |
| const | -417.1743 | 328.717 | -1.269 | 0.251 | -1221.516 to 387.168 |
| Germany | 2.9544 | 0.699 | 4.229 | 0.006 | 1.245 to 4.664 |
| Russia | 2.2726 | 1.021 | 2.226 | 0.068 | -0.226 to 4.771 |
| France | 6.9451 | 2.336 | 2.973 | 0.025 | 1.228 to 12.662 |
| Portugal | -6.3903 | 7.537 | -0.848 | 0.429 | -24.832 to 12.051 |

Table 5. Coefficients

|  |  |
| --- | --- |
| **Test** | **Value** |
| Omnibus | 2.047 |
| Prob(Omnibus) | 0.359 |
| Jarque-Bera (JB) | 0.860 |
| Prob(JB) | 0.650 |
| Skew | -0.026 |
| Kurtosis | 1.631 |
| Durbin-Watson | 1.577 |
| Cond. No. | 1.35e+04 |

Table 6. Residual Tests

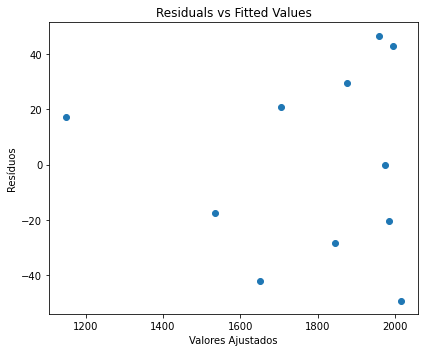


Figure 9. Residuals vs Fitted Values

What we can interpret from this specific plot is:

* **Pattern:** There appears to be a slight pattern where the residuals are not randomly scattered, particularly towards the higher end of fitted values, indicating that the model does not capture all the systematic information in the data.
* **Potential Heteroscedasticity:** The spread of residuals seems to increase with the fitted values. This fan or cone-shaped pattern (widening spread as we move to the right) suggests heteroscedasticity, meaning the variance of the error terms is not constant.
* **Outliers:** The plot shows some points with a very high or low residual value. These points may be outliers in the data, which can have a high leverage effect on the linear regression line and potentially distort the overall model fit.

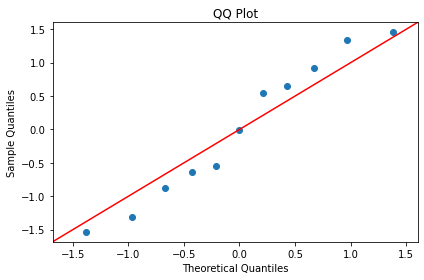


Figure 10. QQ Plot

The QQ plot, on the other hand, is designed to illustrate how well the residuals match a normal distribution.

* **Central Tendency:** The majority of points cluster around the red line, particularly in the center of the plot. This suggests that the residuals are normally distributed around the mean.
* **Tail Behavior:** There is a slight deviation from the line in the upper tail (top right of the plot), indicating that the residuals have some extreme values that are greater than what you would expect in a normal distribution. This can be a sign of outliers or "heavy tails" in the distribution of residuals.
* **Overall Distribution Shape:** Despite the slight deviation in the upper tail, there isn't a systematic deviation from the red line. This means that the residuals do not exhibit a strong skewness. If the residuals were skewed, we would expect to see a systematic curve above or below the line.

|  |  |
| --- | --- |
| **Variable** | **VIF** |
| const | 627.779285 |
| Germany | 7.039836 |
| Russia | 3.252379 |
| France | 4.908781 |
| Portugal | 3.800328 |

Table 7. Variance Inflation Factor

Furthermore, we investigated multicollinearity through the Variance Inflation Factor (VIF). While a VIF **above 10** is often considered a sign of serious multicollinearity, our model exhibited moderate VIFs for Germany, France, and Portugal, suggesting that while there might be some interdependency among predictors, **it wasn't severe enough to undermine the model's validity.**

Despite the insights afforded by our model, which achieved an **R-squared of 0.984**, indicating a high proportion of variance in Europe's emissions could be explained by our independent variables, we recognized limitations. The coefficient for Portugal, notably, had a large p-value and a negative sign, which could imply overfitting or that Portugal's emissions trend during the even years might diverge from the overall European trend.

As for the model's predictive capability, the lack of data for the year 2015 prevented a forecast. Nonetheless, the high R-squared and the model's overall significance as indicated by the F-statistic (93.23) and its associated p-value (1.57e-05) suggested the model could potentially be a robust tool for understanding Europe's emissions landscape.

#### Acknowledgment

I would like to extend my sincerest gratitude to all those who have supported me during this project. My journey has been significantly enriched and guided by the invaluable contributions of my friends and family, whose unwavering support and encouragement have been the backbone of my academic pursuits.

Special thanks are owed to my theoretical professor, Ana Maria Dias Madureira Pereira, whose profound insights and guidance have been instrumental in shaping my understanding and approach towards the theoretical aspects of this study. Her dedication to excellence and scholarly rigor has inspired me to delve deeper into my research, fostering a profound appreciation for the intricacies of our subject matter.

Equally, I am immensely grateful to my practical professor, Teresa Paula Soares De Araujo, whose expertise and practical teachings have immensely contributed to my development and proficiency in the practical applications of our discipline. Her hands-on approach and commitment to fostering a deep understanding of practical methodologies have been crucial in translating theoretical knowledge into practical skills.

Their collective wisdom and mentorship have been pivotal in my academic growth and development, guiding me through challenges and encouraging me to strive for excellence. It is with heartfelt appreciation that I acknowledge their significant impact on my journey.

# Conclusion

This paper has provided an in-depth statistical analysis of CO2 emissions from Portugal and selected regions worldwide, spanning from 1900 to 2021. Through a combination of exploratory data analysis, statistical inference, correlation, and regression modeling, significant insights into the trends, sources, and impacts of CO2 emissions have been uncovered.

**Historical Trends**: The analysis revealed a clear pattern of increasing CO2 emissions over the century, with notable peaks in Portugal in 2005 and various trends in sector-specific emissions. The early to mid-20th century showed relatively stable and low levels of emissions, indicative of minimal industrial activity. A significant increase was observed from the 1950s onwards, peaking towards the late 20th and early 21st centuries, reflecting increased industrialization, reliance on fossil fuels, and higher energy consumption due to economic development and population growth​​.

**Sector-Specific Emissions**: The report detailed the contributions of different sectors to CO2 emissions in Portugal, highlighting significant emissions from cement production, coal combustion, gas, and oil. This underscores the importance of targeted interventions in these sectors to mitigate environmental impact​​.

**Comparative Analysis**: Comparative analyses of CO2 emissions per capita between Portugal and Spain, and among major global emitters (the United States, China, India, the EU, and Russia), underscored the varied pace of industrialization, energy consumption growth, and the effectiveness of environmental policies across regions. Notably, China emerged as a dominant emitter, particularly in coal CO2 emissions, highlighting its heavy dependency on coal for energy​​.

**Statistical Inferences and Predictions**: The application of statistical inference and regression modeling allowed for a nuanced understanding of the factors influencing CO2 emissions. These models have provided a foundation for predicting future emissions trends, underscoring the influence of economic activities, energy consumption patterns, and policy measures on regional and global CO2 emission trajectories​​.

##### References

|  |  |  |
| --- | --- | --- |
| [1] |  | NumFOCUS, Inc., “pandas.pydata,” pandas, [Online]. Available: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.idxmax.html. [Acedido em 21 03 2024]. |
| [2] |  | Instituto Superior de Engenharia do Porto, “moodle.isep.ipp.pt,” [Online]. Available: https://moodle.isep.ipp.pt/course/view.php?id=4965. [Acedido em 31 03 2024]. |
| [3] |  | NumFOCUS, “matplotlib,” matplotlib, [Online]. Available: https://matplotlib.org/. [Acedido em 2024 03 31]. |
| [4] |  | A. . Beevor, The Battle for Spain: The Spanish Civil War 1936–1939, ed., vol. , , : Weidenfeld & Nicolson, 2006, p. 7. |
| [5] |  | T. . Hastie, R. . Tibshirani e J. . Friedman, The Elements of Statistical Learning, ed., vol. , , : , 2009, p. . |
| [6] |  | NumPy Developers, “numpy,” [Online]. Available: https://numpy.org/doc/stable/. [Acedido em 02 04 2024]. |
| [7] |  | The SciPy community, “scipy,” [Online]. Available: https://docs.scipy.org/doc/scipy/. [Acedido em 02 04 2024]. |
| [8] |  | statsmodels, “statsmodels.stats.multicomp.pairwise\_tukeyhsd¶,” [Online]. Available: https://www.statsmodels.org/dev/generated/statsmodels.stats.multicomp.pairwise\_tukeyhsd.html. [Acedido em 03 04 2024]. |
| [9] |  | M. Waskom, “seaborn,” [Online]. Available: https://seaborn.pydata.org/. [Acedido em 03 04 2024]. |