Exploração e análise dos dados de emissões de CO2

Pedro Allen
LEI-DEI
ISEP
Senhora da Hora, Portugal
1211266@isep.ipp.pt

Paulo Reis
LEI-DEI
ISEP
Vila Nova de Gaia, Portugal
1081376@isep.ipp.pt

Rita Azevedo
MEI-DEI
ISEP
Rio Tinto, Portugal
1231439@isep.ipp.pt

***Abstract*— As emissões de CO2 ocorrem diariamente em todo o mundo, e a produção de dióxido de carbono é proveniente de várias fontes como da queima, petróleo e outros que vão ser abordados. A produção de CO2 por país pode mudar ao longo do tempo e a densidade populacional é um dos fatores. É recolhido dados CO2 de diferentes países, juntamente com outros dados relevantes, para realizar análises exploratórias, inferências estatísticas, correlações e análises de regressão.**

Keywords*— *CO2, analise, correlação, regressão, inferência, estatísticas

I. Introdução

As atividades industriais têm contribuído para o aumento das emissões de carbono na atmosfera devido aos crescentes níveis de industrialização e urbanização em muitos países em desenvolvimento. Isso também resultou num aumento significativo na concentração atmosférica global de gases de efeito estufa antropogénicos, incluindo CO2, o que consequentemente tem impulsionado o aquecimento global e as alterações climáticas. Além disso, a desflorestação exemplifica a degradação ambiental, deixando impactos devastadores no nosso planeta. A contínua derrubada de árvores diminui o nosso fornecimento de oxigénio e reduz a absorção de CO2 pelas plantas.[1]

Foi analisado um conjunto de dados que inclui emissões de CO2 de diferentes fontes, juntamente com emissões de metano e óxido nitroso, de 1900 a 2021 em diversos países e regiões. Este conjunto de dados, combinado com outros dados relevantes, foi explorado e analisado para extrair insights e conclusões mais concretas. Estes insights incluem a evolução das emissões de CO2 ao longo do tempo com outras análises relevantes. Além da análise preliminar dos dados, a nossa abordagem abrange inferências estatísticas feitas através da análise de amostras aleatórias. Envolvemo-nos em estudos de correlação detalhados e aplicamos modelos de regressão linear para investigar e interpretar as associações entre variáveis. Este processo analítico permite-nos tirar conclusões robustas e informadas do conjunto de dados que estamos a avaliar. Neste estudo utilizamos a linguagem de programação Python e a ferramenta Jupyter Notebook na framework Anaconda.

II. Objetivos

Os objetivos principais do trabalho incluem uma análise e exploração detalhadas dos dados fornecidos, visando extrair conclusões científicas sobre as emissões de CO2 ao longo do tempo. Além disso, serão conduzidos estudos matemáticos mais aprofundados, envolvendo análise inferencial, correlação e regressão, para possibilitar a extração de outras conclusões relevantes dos dados em estudo.

III. Organização do Artigo

Para facilitar a leitura do artigo, a informação é dividida e organizada por capítulos e subcapítulos. O artigo contém um tópico dedicado a definições e formulas matemáticas para ajudar o leitor a melhor compreender o que está a ser abordado.

IV. Análise e Exploração de Dados

O conjunto de dados fornecido para estudo consiste em vários dados coletados de diferentes países e regiões, permitindo a exploração de diferentes fontes CO2. As informações contidas no conjunto de dados estão resumidas na tabela abaixo com a respetiva designação.

TABLE I. Dados

Nome da Coluna	Designação
country	Região Geográfica
year	Ano de observação
population	População por ano e região
gdp	Produto interno bruto medido em dólares internacionais usando preços de 2011 para ajustar as mudanças de preços ao longo do tempo (inflação) e as diferenças de preços entre países. Calculado multiplicando o PIB per capita pela população.
cement_co2	Emissões anuais de dióxido de carbono (CO2) do cimento, medidas em milhões de toneladas
co2	Emissões totais anuais de dióxido de carbono (CO2), excluindo alterações no uso do solo, medidas em milhões de toneladas.
coal_co2	Emissões anuais de dióxido de carbono (CO2) do carvão, medidas em milhões de toneladas
energy_per_capita	Consumo de energia primária per capita, medido em quilowatts-hora por pessoa por ano.
energy_per_gdp	Consumo de energia primária por unidade de produto interno bruto, medido em quilowatts-hora por dólar internacional.
flaring_co2	Emissões anuais de dióxido de carbono (CO2) provenientes da queima, medidas em milhões de toneladas.
gas_co2	Emissões anuais de dióxido de carbono (CO2) do gás, medidas em milhões de toneladas.
methane	Emissões totais de metano, incluindo alterações no uso do solo e silvicultura - As emissões são medidas em milhões de toneladas de equivalentes de dióxido de carbono.
nitrous_oxide	Emissões totais de óxido nitroso, incluindo alterações no uso do solo e silvicultura - As emissões são medidas em milhões de toneladas de equivalentes de dióxido de carbono.
oil_co2	Emissões anuais de dióxido de carbono (CO2) provenientes do petróleo, medidas em milhões de toneladas.

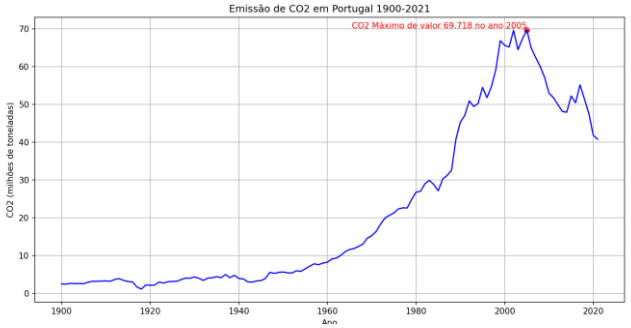

Fig.1 Emissões CO2 em Portugal 1990-2021

Analisamos as emissões de CO2 em Portugal ao longo dos últimos 121 anos, observando um aumento notável desde 1960 conforme representado na figura 1, atingindo o seu pico em 2005, com um valor de 69.718 milhões de toneladas. Houve uma queda acentuada a partir desse ano, com um aumento observado entre 2015 e 2020. O pico em 2005 coincide com o início da utilização de energia renovável em Portugal, ajudando na luta contra as alterações climáticas.[2]

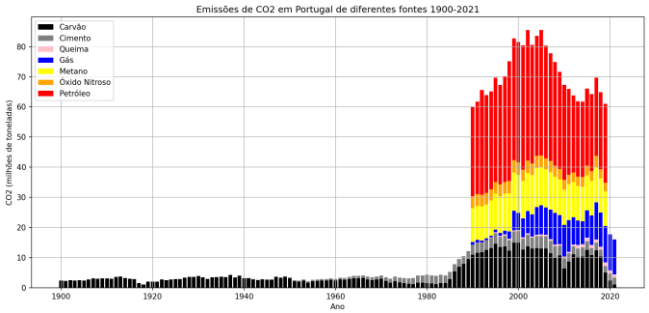

Fig.2 Emissões CO2 em Portugal de diferentes fontes CO2 1990-2021

Excluindo as emissões de carvão e cimento, os dados das emissões provenientes de outras fontes apenas ficaram disponíveis por volta de 1990, enquanto os dados das emissões de carvão estão disponíveis desde aproximadamente 1950. Ao examinar os dados, é evidente que, ao longo dos últimos 30 anos, uma parte significativa das emissões tem origem no petróleo, enquanto a proporção mais pequena provém da queima. Além disso, tanto as emissões provenientes da queima como as emissões de gás têm mostrado uma tendência crescente ao longo dos anos. Observando também que o conjunto de dados não contém informações sobre as emissões de metano, óxido nitroso e petróleo para os anos de 2020 e 2021. O elevado consumo de petróleo em Portugal deve-se principalmente a veículos como carros, segundo o ChatGPT, bem como refinarias, transporte marítimo e o uso de óleo combustível para aquecimento. Não é possível assumir que o consumo de petróleo foi o mais elevado em Portugal antes de 1990 porque esses dados não são conhecidos.

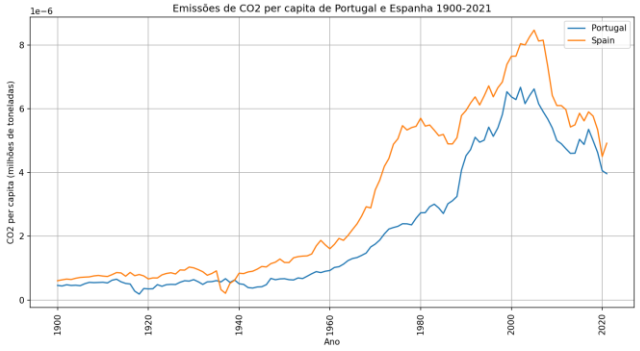

Fig. 3 Emissões CO2 per capita em Portugal e Espanha 1900-2021

No gráfico acima, podemos observar as emissões per capita de Portugal e Espanha ao longo do tempo, desde o ano de 1900 até 2021. Espanha consistentemente apresenta valores mais elevados em comparação com Portugal. Não conseguimos encontrar informações específicas para justificar esta discrepância, mas o ChatGPT sugere que pode estar relacionada a fatores como um maior consumo de energia per capita, uma economia maior, preferência da população por veículos privados em vez de transporte público em Espanha, e outros fatores contribuintes.

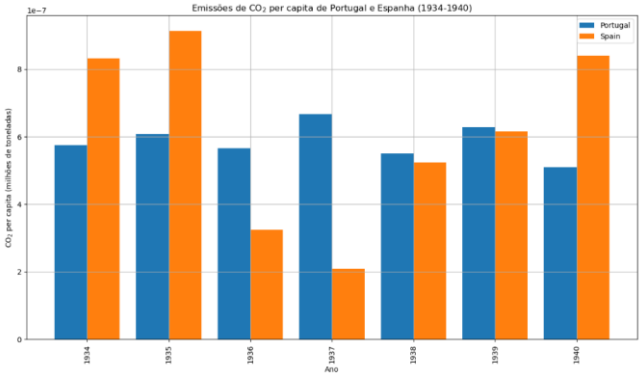

Fig.4 Emissões CO2 em Portugal e Espanha 1934-1940

Para examinar os valores com mais detalhe, criámos um gráfico de barras para mostrar os valores entre os anos de 1934 e 1940, onde as emissões per capita de Portugal excedem as da Espanha de 1936 a 1939. Não conseguimos encontrar uma razão científica específica para justificar dado aos anos que já ocorreu é difícil encontrar artigos, mas de acordo com o ChatGPT, a razão para esta discrepância pode ser atribuída a fatores como diferenças no desenvolvimento industrial, estabilidade económica ou variações nos padrões de consumo de energia entre os dois países durante esse período.

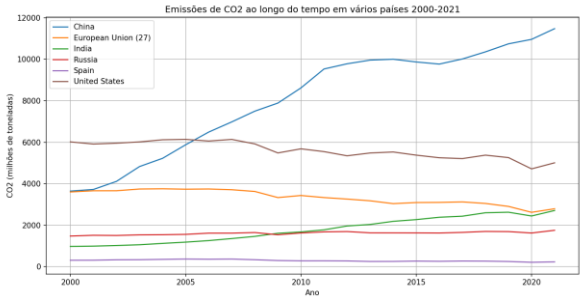

Fig.5 Emissões CO2 ao longo do tempo em diferentes países 2000-2021

Comparamos as emissões de CO2 ao longo do tempo dos países China, União Europeia 27, Índia, Rússia, Espanha e Estados Unidos ao longo dos últimos 21 anos. Durante este período, a China teve um aumento de emissões CO2, enquanto os outros países permaneceram mais constantes ao longo dos últimos 21 anos. Este aumento pode dever-se ao crescimento industrial chinesa. [3]

TABLE II. MÉDIA DAS EMISSÕES CO2 POR FONTE E POR PAÍS 2000-2021

Média das emissões do cimento em milhões de toneladas	
China	599.141
União Europeia (27)	81.488
India	91.512
Rússia	21.837
Espanha	11.972
Estados Unidos	40.055
Média das emissões do carvão em milhões de toneladas	
China	5920.797
União Europeia (27)	1049.236
India	1123.795
Rússia	413.504
Espanha	59.527
Estados Unidos	1750.037
Média das emissões proveniente da Queima em milhões de toneladas	
China	1.722
União Europeia (27)	21.132
India	2.661
Rússia	43.061
Espanha	2.925
Estados Unidos	52.728
Média de emissões de gás em milhões de toneladas	
China	287.021
União Europeia (27)	774.871
India	92.464
Rússia	766.698
Espanha	62.223
Estados Unidos	1364.198
Média das emissões do metano em milhões de toneladas	
China	1015.726
União Europeia (27)	407.444
India	617.360
Rússia	599.007
Espanha	39.032
Estados Unidos	639.154
Média das emissões do óxido nitroso em milhões de toneladas	
China	476.530
União Europeia (27)	238.482
India	228.242
Rússia	58.484
Espanha	21.342
Estados Unidos	259.030
Média das emissões do petróleo em milhões de toneladas	
China	1116.257
União Europeia (27)	1374.161
India	469.662
Rússia	353.289
Espanha	156.162
Estados Unidos	2379.692

Na tabela acima, temos as emissões de CO2 por fonte, medidas em milhões de toneladas, para os países China, União Europeia (27), Índia, Rússia, Espanha e Estados Unidos entre os anos de 2000 e 2021. Ao analisar o conjunto de dados antes de construir a tabela, notamos que alguns países têm valores nulos em diferentes anos relacionados com as emissões de metano e óxido nitroso.

Observando os dados fornecidos, podemos ver que a China produz as emissões mais elevadas de cimento, carvão, metano e óxido nitroso, enquanto os Estados Unidos produzem a maioria das emissões provenientes da queima, gás e petróleo. Por outro lado, a Espanha emite menos emissões de cimento, carvão, gás, óxido nitroso e petróleo, com a China emitindo menos emissões provenientes da queima. Como alguns dados para metano e óxido nitroso estão em falta, esses valores podem mudar se tivéssemos os valores em falta.

V. DEFINIÇÕES E FORMULAS MATEMÁTICAS

A. Nível de significância

O nível de significância α é a probabilidade de ocorrer um erro do tipo I.

$\alpha = P(H1|H0)$

B. Amostras emparelhadas

Amostras emparelhadas, são amostras que teem relação uma com a outra.

C. H0 (hipótese nula)

H(0) é a hipótese nula, não há diferença (ou seja, μ_diferença = 0).

D. H1 (hipótese alternativa)

Existe uma diferença (ou seja, μ diferença $\neq$ 0).

E. Verificação das suposições

Assegurar que os dados não violam as suposições do teste utilizado.

F. *Cálculo das diferenças*

Calcular a diferença de cada par de observações (por exemplo, GDP de Portugal menos GDP da Hungria para cada ano na amostra).

G. *Decisão*

Se o valor p for menor que 0,05, rejeitaríamos a hipótese nula, concluindo que há uma diferença significativa entre as médias das amostras emparelhadas.

VI. Inferência Estatistica

A análise inferencial representa uma abordagem de raciocínio indutivo com o objetivo de extrapolar conclusões de um subconjunto específico designado de amostra, para um contexto mais amplo, população. Ao estabelecer uma hipótese, o procedimento envolve a escolha de um modelo estatístico que se alinhe com o processo de geração de dados, o que forma a base para inferências subsequentes. Este método utiliza características observadas na amostra para fazer generalizações sobre a população maior. Para todos os testes de hipótese efetuados considerou-se um grau de significância de 5%.

A. *4.2.1 Teste á média GDP de Hungria e Portugal numa amostra de 26*

1) *Análise e amostra de dados*

Fig. 6 Média do PIB para Portugal e Hungria (de 1900-2021).

Após a leitura e tratamento dos valores nulos apresentados em cada país, realizou-se o cálculo da média amostral de cada país que pode ser observado na tabela 3, concluindo que a média amostrar em Portugal é maior que a da Hungria.

TABLE III. Média Produto interno Bruto(GDP) por país

País	Average GDP
Portugal	1.0234e+5 millions $
Hungria	4.2188e+4 millions $

2) *Processo de inferencia estatistica*

Nesta situação, dado que que foram removidos os valores nulos de cada um dos países em 26 amostras, identifica-se que se trata de amostras emparelhadas. Nesta situação o teste mais adequado é o teste t, utilizando-o relativamente à média de GDP de cada país.

3) *Formulação das hipoteses*

$$H_0 : \bar{X}_{PT} \geq \bar{X}_{HU} \quad Vs \quad H_0 : \bar{X}_{PT} < \bar{X}_{HU}$$

Fig. 7 Formulação das hipóteses

Na hipótese nula (H0) afirma-se que a média da população de Portugal (PT) é maior ou igual à média da população da Hungria (HU). Na hipótese alternativa (H1) sugere-se que a média da população de Portugal é menor que a média da população da Hungria.

Após realizado o teste, obteve-se o valor p-value de 0.0076. No contexto do nível de significância estabelecido a 5%, o valor p-value obtido, permite-nos concluir que há uma diferença estatisticamente significativa entre as médias em estudo. Com este valor p substancialmente inferior ao limiar de 0,05, rejeita-se a hipótese nula.

B. *4.2.2 Teste á média GDP de Hungria e Portugal numa amostra de 12*

1) *Análise e amostra de dados*

Nesta situação, as amostras são identificadas como independentes, sendo o mais adequado o teste t. No entanto, para utilizar este teste, é necessário verificar algumas suposições.

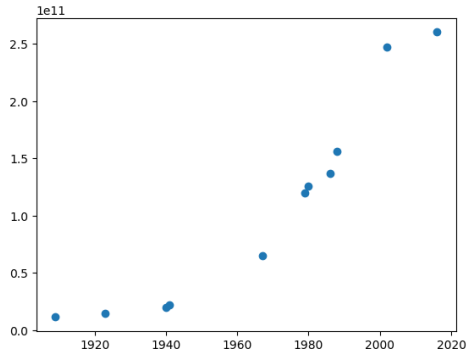

Fig. 8 Média do GDP de Portugal 1900-2021

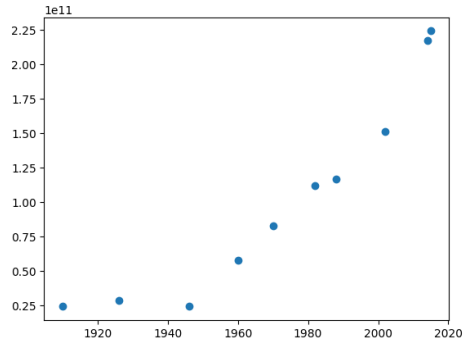

Fig. 9 Média do GDP da Hungria 1900-2021

2) *Suposições*

As amostras são consideradas independentes, conforme garantido pelo método de amostragem. Para verificar a normalidade das amostras, foi realizado o teste de Shapiro, onde testamos a hipótese nula de que as populações fornecidas

seguem uma distribuição normal. Os valores de p-value obtidos foram 0.115 e 0.118 para Portugal e Hungria, respetivamente. Como ambos os valores de p-value são maiores que o nível de significância de 0.05, não podemos rejeitar a hipótese nula, sugerindo que os dados seguem uma distribuição normal.

Para a homogeneidade das variâncias, foi realizado o teste de Levene para verificar se em ambos os casos são aproximadamente iguais, obtendo-se um valor de p-value de 0.567. Como este valor é maior que o nível de significância considerado, assumimos que as variâncias são iguais, não rejeitando a hipótese nula.

3) Formulação das hipoteses

$$H_0 : \bar{X}_{PT} \geq \bar{X}_{HU} \quad Vs \quad H_1 : \bar{X}_{PT} < \bar{X}_{HU}$$

Fig. 10 Hipótese da média do GDP de Portugal ser superior á da Hungria 1900-2021

4) Processo de inferência estatística

Na hipótese nula (H0) afirma-se que a média população de Portugal é maior ou igual à média da população da Hungria (HU). Na hipótese alternativa (H1) sugere-se que a média da população de Portugal é menor que a média da população da Hungria.

Após realizado o teste, obteve-se o valor p-value de 0.937. No contexto do nível de significância estabelecido a 5%, o valor p-value obtido permite-nos concluir que não há uma diferença estatisticamente significativa entre as médias em estudo. Com este valor p superior ao limiar de 0,05, não se rejeita a Hipótese Nula, podendo inferir que a média de GDP de Portugal é superior ao GDP da Hungria nas amostras consideradas.

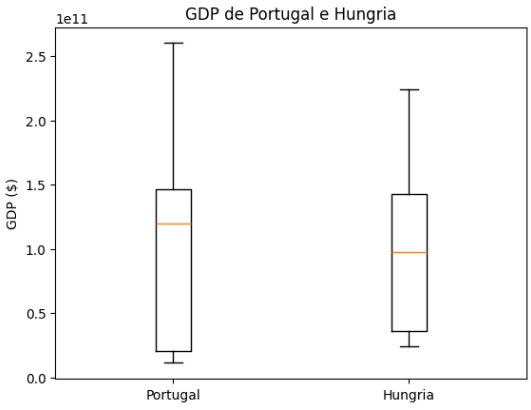

Fig. 11 Boxplot do GDP de Portugal e Hungria

C. 4.2.3 Teste á média GDP da União Europeia, Estados Unidos, Russia , China, India

1) Análise e amostra de dados

Neste caso, pretende-se verificar se há diferenças significativas nas emissões totais de CO2 nas diferentes áreas geográficas. Identificou-se que as amostras são emparelhadas, no entanto existem outliers significativos, não se reunindo assim as condições necessárias para a utilização do teste paramétrico One-Way ANOVA. Em alternativa usamos o teste não paramétrico de Friedman.

2) Formulação das hipoteses

$$H_0 : \mu_{US} = \mu_{RU} = \mu_{CN} = \mu_{IN} = \mu_{EU} \quad Vs \quad H_1 : \exists \quad i, j : \mu_i \neq \mu_j$$

Fig. 12 Hipótese da igualdade das variâncias

3) Processo de inferência estatística

Após a realização do teste, obteve-se o valor p-value de 3.7250e-07. No contexto do nível de significância estabelecido a 5%, o valor p-value obtido permite-nos concluir que há uma diferença estatisticamente significativa entre as médias em estudo. Com este valor p muito baixo, próximo de 0, relativamente ao limiar de 0,05, rejeita a Hipótese Nula podendo inferir que que as emissões de CO2 são diferentes.

VII. Correlação e Regressão

A. 4.3.1 Correlação entre os diferentes continentes

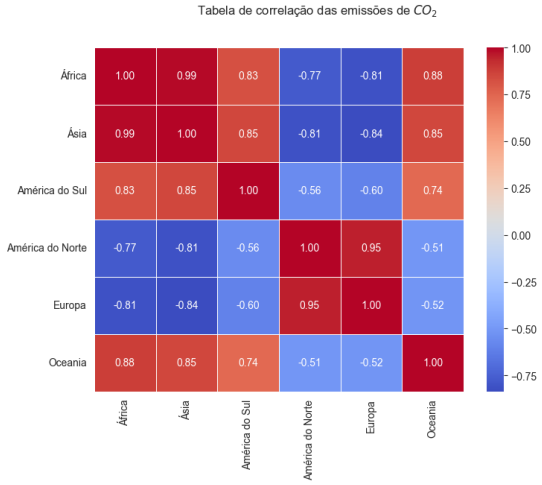

Fig. 13 Tabela de correlação das emissões de CO2

Existe uma correlação positiva muito entre os valores das emissões de África e da Ásia. Verifica-se também uma correlação positiva noutro par de zonas: América do Norte e Europa. Estas correlações são tão fortes que o comportamento com outras zonas é similar.
Observa-se também que existem correlações negativas entre as zonas de cada um daqueles grupos, ligeiramente mais forte no caso da Asia e Europa.

B. *4.3.2 Análise da correlação e regressão nos diferentes continentes*

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_4+\epsilon \quad (1)$$

$$Y=-1653{,}37-0{,}58.X_{Germany}+2{,}09.X_{Russia}+12{,}74.X_{France}+1{,}27.X_{Portugal}+\epsilon \quad (2)$$

Para encontrar um modelo de regressão linear múltipla válido para os dados que temos, devemos não só determinar o modelo, mas também validar o mesmo.

Y - Emissão de CO2 provenientes do carvão nos anos pares do século XXI na Europa ('Europe').

X1 - Emissão de CO2 provenientes do carvão nos anos pares do século XXI na Alemanha.

X2 - Emissão de CO2 proveniente do carvão nos anos pares do século XXI na Rússia.

X3 - Emissão de CO2 proveniente do carvão nos anos pares do século XXI na França.

X4 - Emissão de CO2 proveniente do carvão nos anos pares do século XXI em Portugal.

O modelo será definido por esta variáveis, e poderá ser descrito como se verifica na equação (1). Este modelo, se validado, deverá ser usado para estimar o valor pedido no ponto ii).

Calculado o modelo e os seus valores, verificamos a sua homocedasticidade, independência de resíduos e multicolinearidade de modo a validar o mesmo.

O modelo obtido é o apresentado na equação (2) e apresenta um R^2 ajustado de 98.98%. Obtivemos os valores de p-value do nível de significância de acordo com a tabela IV.

TABLE IV. VALORES P-VALUE DOS NIVEIS DE SIGNIFICÂNCIA

País	p-value
Alemanha	0.442943
Rússia	0.000488
França	0.000233
Portugal	0.826289

Fizemos um gráfico para verificar a homocedasticidade apresentado no gráfico da figura 14, assim analisar os quartis presente no gráfico da figura 15. Fizemos ainda um teste de Shapiro ao valor dos resíduos do modelo obtendo um valor de p-value de 0.571.

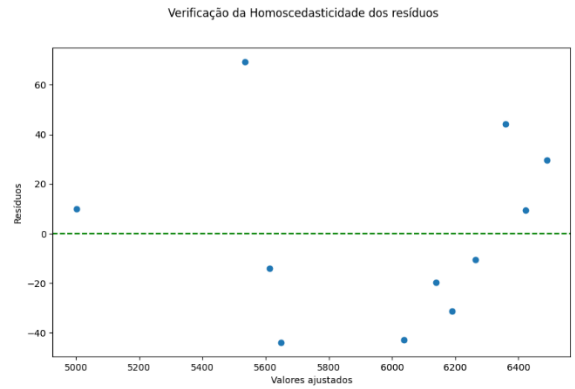

Fig. 14 Verificação das Homoscedasticidade dos resíduos

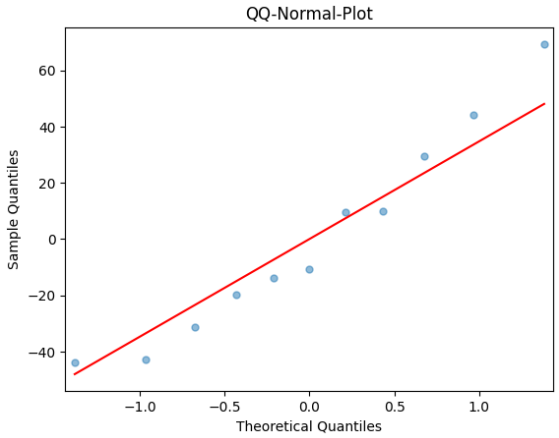

Fig. 15 QQ Normal Plot

Fizemos um teste de Durbin–Watson para verificar a independência dos resíduos obtendo um valor de 2.21. Por último fizemos um diagnóstico da Multicolinariedade usando o fator de inflação de variância (VIF), obtendo os valores apresentados na tabela V.

TABLE V. VALORES VIF

País	VIF
Alemanha	1576.1765
Rússia	140.2838
França	1654.5729
Portugal	414.0853

Com base no nosso modelo, a previsão para a emissão de CO2 na Europa em 2015 é de 5591.926 milhões de toneladas.

VIII. CONCLUSÃO

Na fase de análise e exploração, podemos concluir que a análise dos dados de emissões de CO2 ao longo do último século revela tendências e padrões significativos nos níveis de emissões, em específico em Portugal, quando é introduzido as energias renováveis. Embora as razões para certas discrepâncias nas emissões permaneçam elusivas a níveis científicos em uma década e país concreto, o uso de ferramentas de IA como o ChatGPT fornece esclarecimentos

fatores subjacentes que impulsionam essas diferenças, uma vez que alguns desses eventos ocorreram há muitos anos.

Nos vários casos em que aplicamos inferência estatística sobre a amostras aleatórios dos dados, a utilização de metodologias adequadas, permitiram-nos inferir informação não presente nas amostras. Ao longo deste artigo colocamos os resultados junto dos respetivos estudos e valores obtidos nos testes, pelo que as conclusões são acompanhadas dos mesmos. Os dados obtidos pelo nosso modelo foram significativamente próximos do real, consideramos aceitável para poder fazer algumas deduções sobre valores futuros de valores CO2.

Num futuro trabalho poderá fazer-se o tratamento dos dados, preenchendo alguns valores nulos com as técnicas adequadas e realizar mais estudos aos dados.

[1] Kelvin O. Yoro, M.O. Daramola. CO2 emission sources, greenhouse gases, and the global warming effect. 2020

[2] Gonçalves, Júlia Diniz Jacques. Avaliação das Poupanças nas Emissões de CO2 Geradas Pela Produção de Energia Renovável no Sistema Elétrico Português Entre 2005 e 2017. 2018. Instituto Superior de Engenharia do Porto

[3] Ting Gao. Regional Science and Urban Economics. 2004.

[4] Friedman's Two-way Analysis of Variance by Ranks - Analysis of k-Within-Group Data with a Quantitative Response Variable. Available: hcfried.PDF (unl.edu)

[5] Logistic Regression Four Ways with Python. Available: Logistic Regression Four Ways with Python | UVA Library (virginia.edu)

[6] Linear Regression (Python Implementation). Available: Linear Regression (Python Implementation) - GeeksforGeeks

[7] Simple Linear Regression in Python. Available: Simple Linear Regression in Python | by Shuvrajyoti Debroy | Medium

[8] Mastering Linear Regression with Statsmodels. Available: Mastering Linear Regression with Statsmodels | by Luís Fernando Torres | LatinXinAI | Medium