Análise de Dados - Análise Exploratória de Dados

Daniel Braga
Departamento de Engenharia Informática
Instituto Superior de Engenharia do Porto
Porto, Portugal
1200801@isep.ipp.pt

Gonçalo Nogueira
Departamento de Engenharia Informática
Instituto Superior de Engenharia do Porto
Porto, Portugal
1201525@isep.ipp.pt

***Abstract* - Este artigo tem como objetivo a descrição e análise de dados para aplicar no enunciado do primeiro Trabalho Prático da cadeira de Análise de Dados em Informática. De seguida, como Referências teóricas foram descritas as seguintes técnicas: análise exploratória de dados, inferência estatística, correlação e regressão linear, e para cada exercício foram explicados os métodos usados para o resolver e uma breve conclusão. Para este estudo foram utilizadas as capacidades de análise de dados da linguagem Python, usando o Spyder.**

I. Introdução

O presente artigo científico foi desenvolvido no âmbito da unidade curricular Análise de Dados em Informática (ANADI), do 2º semestre do 3ºano da Licenciatura em Engenharia Informática (LEI) do Instituto Superior de Engenharia do Porto (ISEP). Neste âmbito foi proposta a realização de uma análise exploratória de dados, inferência estatística, correlação e regressão linear. Com este artigo deverá ser possível fazer uma análise sobre os dados, através dos conhecimentos adquiridos nesta unidade curricular, usando a linguagem Python e o Spyder.

II. Revisão da Literatura

A. *Análise exploratória de dados*

A análise exploratória de dados (AED) é uma etapa inicial essencial na ciência de dados. Envolve a sumarização e visualização dos dados para compreender sua estrutura, detectar padrões e anomalias, e formular hipóteses iniciais. Durante a AED, técnicas como sumarização estatística, detecção de outliers, exploração de relacionamentos entre variáveis, e análise de distribuição são aplicadas. Esses insights iniciais orientam as etapas subsequentes do projeto de ciência de dados, ajudando os cientistas de dados a selecionar as técnicas de modelagem adequadas e extrair conhecimento útil para tomada de decisão..

B. *Testes paramétricos*

Os testes de hipóteses investigam se uma afirmação sobre um parâmetro de uma população é verdadeira ou falsa, com base em amostras aleatórias. Por exemplo, determinar se um partido político receberá mais de 30% dos votos em uma eleição.

Ao realizar um teste de hipóteses, consideramos uma hipótese nula (H0) e uma hipótese alternativa (H1). Dependendo da relação entre o valor do parâmetro e um valor fixo (θ0), temos três tipos de testes: bilateral, unilateral à direita e unilateral à esquerda. Podemos cometer dois tipos de erros em um teste de hipóteses: erro do Tipo I (rejeitar H0 quando é verdadeira) e erro do Tipo II (não rejeitar H0 quando é falsa). O nível de significância α é a probabilidade de cometer um erro do Tipo I.

Para realizar um teste paramétrico, seguimos estas etapas: definimos as suposições sobre as distribuições das variáveis, formulamos H0 e H1, construímos uma estatística de teste com distribuição conhecida sob H0, calculamos o valor observado da estatística de teste com base nos dados da amostra e decidimos se rejeitamos ou não H0 com base no valor de p ou no p-value.

C. *Testes não paramétricos*

Os testes não paramétricos são uma classe de testes estatísticos que não requerem a suposição de uma distribuição específica para os dados. Isso os torna úteis quando os dados não seguem uma distribuição conhecida ou quando os pressupostos dos testes paramétricos não são atendidos. Em vez de utilizar parâmetros específicos da distribuição, esses testes baseiam-se em estatísticas de ordem ou ranks dos dados, tornando-os mais robustos em certas situações.

Existem diferentes tipos de testes não paramétricos, cada um adequado para diferentes tipos de análises. Por exemplo, os testes de localização são usados para comparar a mediana de uma população ou para determinar se duas amostras independentes têm medianas diferentes. Exemplos incluem o Teste de Sinais, o Teste de Wilcoxon e o Teste de Mann-Whitney-U.

Os testes não paramétricos oferecem uma alternativa robusta quando os pressupostos dos testes paramétricos não são atendidos ou quando os dados não seguem uma distribuição conhecida. A escolha do teste adequado depende das características dos dados e do tipo de análise que se deseja realizar.

D. *Correlação*

Os testes de correlação são métodos estatísticos utilizados para avaliar a relação entre variáveis. Eles são fundamentais para entender como as mudanças em uma variável estão associadas às mudanças em outra variável. Existem diferentes tipos de testes de correlação, cada um adequado para diferentes cenários de análise.

O Teste de Correlação Linear de Pearson é utilizado quando as variáveis são contínuas e há uma suposição de

relação linear entre elas. O coeficiente de correlação Pearson (r) é calculado para medir a força e a direção dessa relação. Ele varia de -1 a 1, onde valores próximos de 1 indicam uma correlação positiva forte, valores próximos de -1 indicam uma correlação negativa forte e valores próximos de 0 indicam uma correlação fraca.

O Teste de Correlação Ordinal de Spearman é aplicado quando as variáveis são ordinais ou uma é contínua e a outra é ordinal. Ele utiliza o coeficiente de correlação de Spearman (ρ) para avaliar a relação monotônica entre as variáveis. Este teste é uma alternativa robusta ao teste de Pearson em situações onde a relação entre as variáveis não é linear.

Já o Teste de Correlação Ordinal de Kendall é outra alternativa ao teste de Spearman, especialmente útil em amostras pequenas ou com muitos empates. Ele baseia-se no número de pares concordantes e discordantes entre as observações para calcular o coeficiente de correlação de Kendall (τ), que também indica a força e a direção da relação entre as variáveis.

Em resumo, os testes de correlação são ferramentas estatísticas importantes para investigar a associação entre variáveis em um conjunto de dados. A escolha do teste adequado depende das características das variáveis e do tipo de relação que se deseja analisar.

E. Regressão linear

A análise de regressão é usada para modelar a relação entre uma variável dependente Y e uma ou mais variáveis independentes X1, ..., Xp. Quando há apenas uma variável independente, é uma regressão simples; quando há mais de uma, é uma regressão múltipla.

As variáveis podem ser contínuas, discretas ou categóricas. O modelo de regressão linear simples assume uma relação linear entre Y e X, representada por $Y = \beta0 + \beta1X + \varepsilon$, onde os coeficientes β0 e β1 são os coeficientes de regressão e ε representa os erros.

Para validar o modelo, são verificados pressupostos sobre a distribuição dos erros e a independência dos valores observados da variável dependente. A estimação dos coeficientes é realizada minimizando a norma dos erros. O coeficiente de determinação (R^2) é usado para medir o poder explicativo do modelo.

Testes de hipóteses, como o teste t e o teste F, são utilizados para determinar a significância dos coeficientes e do modelo como um todo. A inferência estatística também envolve a verificação de condições sobre os resíduos, como normalidade, homocedasticidade e independência. O diagnóstico de multicolinearidade é feito usando o fator de inflação de variância (VIF). A inclusão de variáveis dummy é necessária para modelar variáveis categóricas.

III. REALIZAÇÃO DO TRABALHO

4.1 Análise e exploração de dados

1) Inicialmente, os dados de CO2 foram carregados a partir do ficheiro “CO_data.csv”. Em seguida, os dados foram filtrados para incluir apenas as informações referentes a Portugal. Utilizando esses dados filtrados, foi criado um gráfico que representa a evolução das emissões de CO2 ao longo do tempo, destacando os anos em que as emissões foram registradas. O gráfico apresenta o ano no eixo x e as emissões de CO2 em milhões de toneladas no eixo y. O ponto máximo no gráfico indica o ano com o maior valor de emissões de CO2.

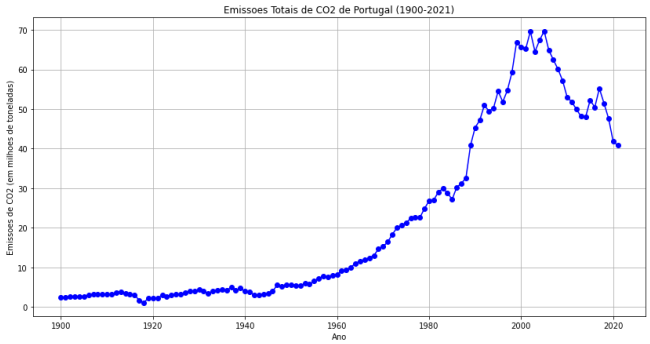

Figura 1 – Gráfico do exercício 4.1.1

2) Inicialmente, os dados foram carregados e filtrados para incluir apenas as informações pertinentes a Portugal.

Em seguida, foram selecionadas várias fontes de emissões, como "cement_co2", "coal_co2", "flaring_co2", "gas_co2", "methane", "nitrous_oxide" e "oil_co2", e os dados de emissões correspondentes a cada fonte foram refletidos num gráfico.

Este gráfico mostra a evolução das emissões de CO2 provenientes de diferentes fontes ao longo do tempo, permitindo uma comparação visual entre elas. Cada linha no gráfico representa uma fonte específica de emissão, com o ano no eixo x e as emissões de CO2 em milhões de toneladas no eixo y. A legenda do gráfico identifica cada linha de acordo com a fonte correspondente de emissão.

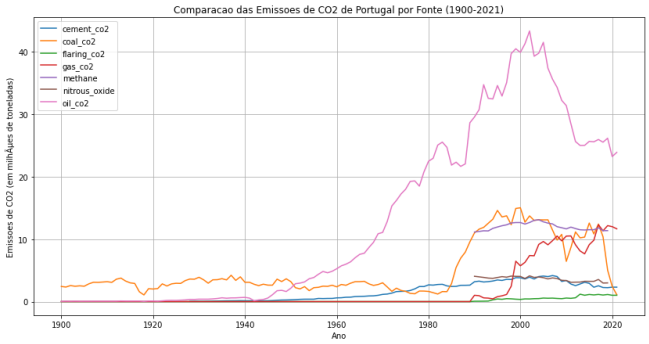

Figura 2 – Gráfico do exercício 4.1.2

3) Após carregar e filtrar os dados para incluir informações específicas de Portugal e Espanha, foram calculadas as emissões de CO2 per capita para cada país. Isso foi feito ao dividir as emissões totais de CO2 pelo número de

habitantes de cada país e multiplicando o resultado por um milhão para obter as emissões em toneladas por pessoa.

Em seguida, os dados das emissões per capita de CO2 foram representados num gráfico, com o eixo x a representar o ano e o eixo y a representar as emissões per capita em toneladas por pessoa. A legenda identifica cada linha com o país correspondente.

Este gráfico proporciona uma visão clara das tendências das emissões de CO2 per capita em Portugal e na Espanha ao longo do período analisado, permitindo uma comparação das trajetórias de ambas as nações em termos de emissões individuais per capita.

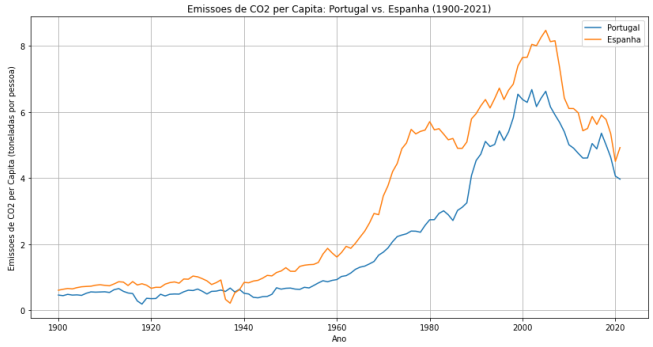

Figura 3 – Gráfico do exercício 4.1.3

4) Primeiramente, os dados foram filtrados para incluir apenas as informações relevantes para as regiões de interesse (Estados Unidos, China, Índia, UE-27 e Rússia) entre 2000 e 2021. Em seguida, as emissões de CO2 provenientes do carvão para cada região foram representadas num gráfico. O eixo x representa o ano, enquanto o eixo y representa as emissões de CO2 em milhões de toneladas.

Para cada região corresponde uma linha no gráfico, destacando as emissões de CO2 provenientes do carvão ao longo do período analisado. Cada linha é identificada com o nome da região correspondente na legenda do gráfico.

Esta visualização permite uma comparação direta das tendências das emissões de CO2 do carvão entre as diferentes regiões ao longo do tempo, destacando variações e padrões de comportamento específicos de cada área geográfica.

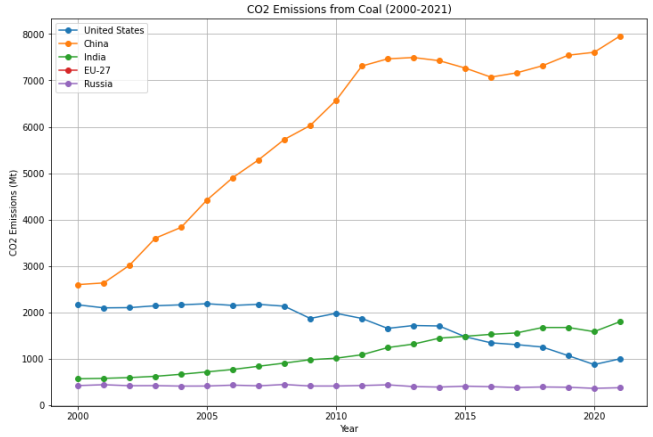

Figura 4 – Gráfico do exercício 4.1.4

5) Inicialmente, os dados foram filtrados para incluir apenas as mesmas regiões do exercício anterior para o mesmo intervalo de tempo. Em seguida, as médias das emissões de CO2 foram calculadas para cada fonte de emissão, incluindo emissões de cimento, carvão, gás, metano, óleo, entre outras, para cada região.

Os resultados foram arredondados para três casas decimais. Posteriormente, os resultados foram apresentados numa tabela, onde as linhas representam as diferentes regiões e as colunas representam as diferentes fontes de emissão de CO2.

Cada célula da tabela contém a média das emissões de CO2 correspondente à região e à fonte de emissão específicas.

	cement_co2	coal_co2	flaring_co2	gas_co2	methane	nitrous_oxide	oil_co2
China	599.141	5920.797	1.722	287.021	1015.726	476.53	1116.257
European Union (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

Figura 5 – Tabea do exercício 4.1.5

4.2 Inferência Estatística

1) Inicialmente, carregamos os dados do CO2 e selecionamos aleatoriamente 30 anos dentro do período de 1900 a 2021 para compor a amostra sampleyears1. Em seguida, filtramos os dados para obter o PIB de Portugal e da Hungria correspondentes aos anos presentes na amostra.

Após a preparação dos dados, realizamos o teste t de duas amostras utilizando a função ttest_ind da biblioteca SciPy. Configuramos a alternativa como 'greater', pois estávamos interessados em verificar se a média do PIB de Portugal era estatisticamente maior do que a média do PIB da Hungria.

Os resultados do teste revelaram um valor de t-estatístico aproximadamente igual a 0.181 e um valor p aproximadamente igual a 0.428. Com um nível de significância de 5% ($\alpha = 0.05$), como o valor p é maior do que 0.05, não temos evidências suficientes para rejeitar a hipótese nula. Portanto, não podemos afirmar que a média do PIB de Portugal foi significativamente maior do que a média do PIB da Hungria durante o período analisado, com base nos anos selecionados pela amostra sampleyears1.

Além disso, representamos num gráfico de caixa os PIBs de Portugal e Hungria para visualizar a distribuição dos dados e facilitar a interpretação dos resultados do teste t.

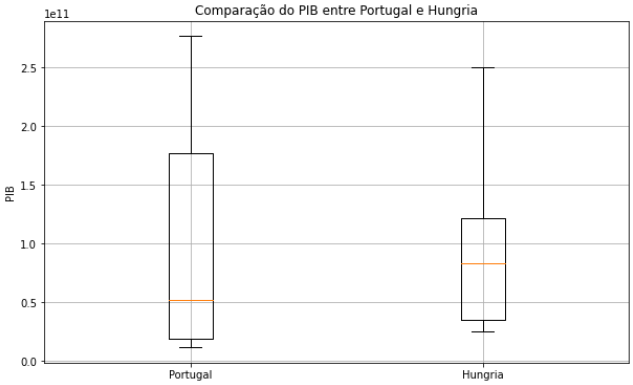

Figura 6 – Gráfico do exercício 4.2.1

2) Para este exercício foi realizado um novo teste t de duas amostras para médias independentes, porém utilizando duas amostras distintas para Portugal e Hungria. As amostras foram selecionadas aleatoriamente para cada país, sendo denominadas sampleyears2 para Portugal e sampleyears3 para Hungria.

Após a definição das amostras, os dados foram filtrados para obter o PIB correspondente aos anos presentes em cada amostra. Em seguida, foi realizado o teste t de duas amostras utilizando a função ttest_ind da biblioteca SciPy, com a alternativa configurada novamente como 'greater'.

Os resultados do teste revelaram um valor de t-estatístico de aproximadamente 0.242 e um valor p de aproximadamente 0.406. Com um nível de significância de 5% ($\alpha = 0.05$), como o valor p é maior do que 0.05, não foram encontradas evidências suficientes para rejeitar a hipótese nula. Portanto, não se pode afirmar que a média do PIB de Portugal foi significativamente maior do que a média do PIB da Hungria durante o período analisado, mesmo utilizando amostras diferentes para cada país.

Além disso, para facilitar a interpretação dos resultados, foi feito um gráfico de barras comparando as médias do PIB de Portugal e Hungria com base nas amostras selecionadas, destacando visualmente as diferenças entre as médias dos dois países.

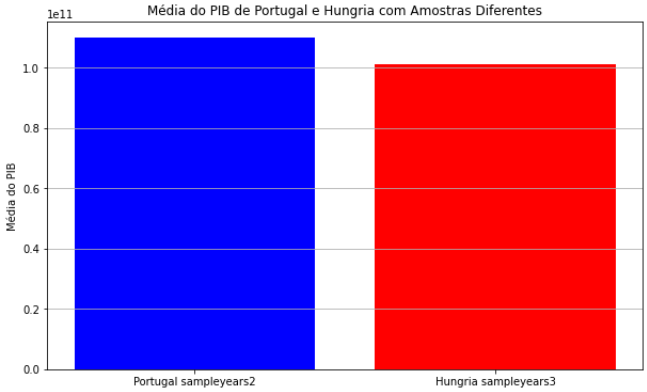

Figura 7 – Gráfico do exercício 4.2.2

3) Primeiramente, os dados foram filtrados para incluir apenas as informações dos países e anos presentes na amostra sampleyears2. Em seguida, as emissões de CO2 para cada país foram extraídas e preparadas para a análise.

Após o preparo dos dados, a ANOVA foi realizada utilizando a função f_oneway da biblioteca SciPy. Essa função calcula a estatística F e o valor p associado, permitindo determinar se há diferenças estatisticamente significativas entre as médias das emissões de CO2 das regiões selecionadas.

Os resultados da ANOVA revelaram que o valor p associado à estatística F é maior que o nível de significância adotado de 5%. Isso indica que não há evidências suficientes para rejeitar a hipótese nula de que as médias das emissões de CO2 das regiões são iguais. Portanto, com base nos anos amostrados, não podemos concluir que há diferenças significativas nas emissões totais de CO2 entre as regiões dos Estados Unidos, Rússia, China, Índia e União Europeia.

Além disso, para auxiliar na interpretação dos resultados, foi feito um gráfico de barras para comparar as médias das emissões de CO2 de cada região. Essa visualização fornece uma representação intuitiva das médias das emissões de CO2 de cada região, permitindo uma comparação direta entre elas e destacando as diferenças.

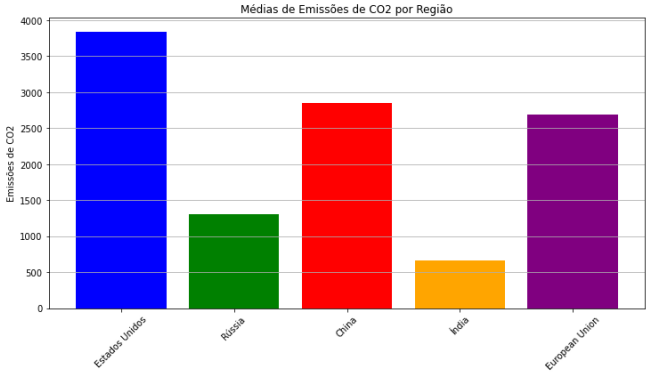

Figura 8 – Gráfico do exercício 4.2.3

4.3 Correlação e Regressão

1) Primeiramente, os dados foram filtrados para incluir apenas as informações das regiões: África, Ásia , América do Sul, América do Norte, Europa e Oceania entre 2000 e 2021. De seguida, as emissões de CO2 provenientes do carvão para cada região foram calculadas e organizadas numa tabela.

Após a preparação dos dados, a tabela de correlação foi calculada usando o método corr() do DataFrame pandas, que calcula o coeficiente de correlação de Pearson entre todas as combinações de pares de regiões. Isso resulta numa matriz de correlação onde cada entrada representa o coeficiente de correlação entre duas regiões.

Para visualizar a tabela de correlação, foi utilizada a biblioteca Seaborn para gerar um mapa de calor (heatmap). Nesse mapa de calor, as células são coloridas de acordo com o valor do coeficiente de correlação, facilitando a identificação de padrões de correlação entre as regiões. Valores mais próximos de 1 indicam uma correlação positiva forte, valores próximos de -1 indicam uma correlação negativa forte e valores próximos de 0 indicam falta de correlação linear.

O resultado final é uma representação visual da correlação entre as emissões de CO2 provenientes do carvão para cada par de regiões, fornecendo insights sobre possíveis relações entre elas ao longo do período analisado.

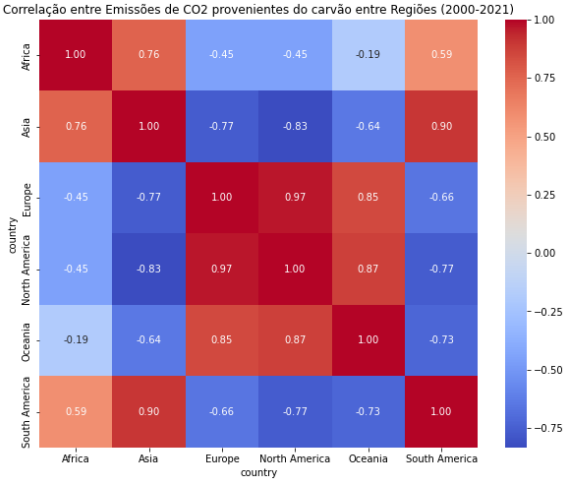

Figura 9 – Tabela do exercício 4.3.1

2)

a) Para construir o modelo de regressão linear, utilizamos dados das emissões de CO2 provenientes do carvão nos anos pares do século XXI na Alemanha, Rússia, França, Portugal e na região europeia. Após ajustar o modelo, encontramos uma equação que nos permite prever as emissões de CO2 na Europa com base nas emissões desses países:

Europe = -417,17 + 2,95*Germany + 2,27*Russia + 6,95*France + -6,39*Portugal

b) Realizamos o teste de Shapiro-Wilk para avaliar a normalidade dos resíduos obtendo os seguintes resultados: Estatística de teste = 0,941 e valor p = 0,527. Como o nosso valor p é maior do que 0,05, podemos afirmar que não existem evidências suficientes para rejeitar a hipótese nula de normalidade dos resíduos, assumindo um nível de significância de 5%. Em outras palavras, os resíduos podem seguir uma distribuição normal.

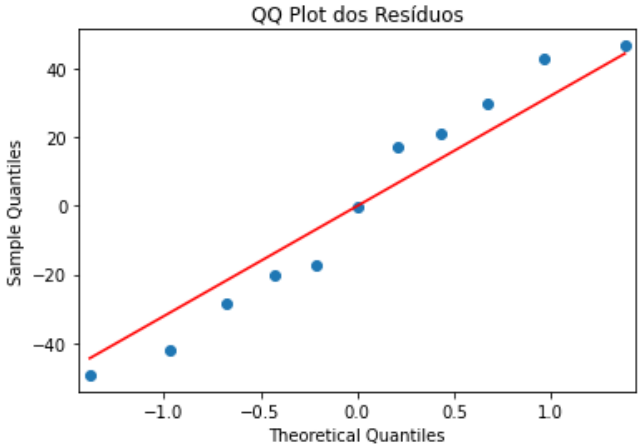

Figura 10 – QQ Plot dos Resíduos

c) Calculamos o fator de inflação da variância (VIF) para cada variável independente:

	feature	VIF
0	const	627.779285
1	Germany	7.039836
2	Russia	3.252379
3	France	4.908781
4	Portugal	3.800328

Figura 11 – VIF para cada variável independente

A Alemanha apresenta um VIF de 7.04 o que indica que há uma moderada multicolinearidade. Seguido da Russia, França e Portugal, com um VIF entre 3.25 e 4.91, o que também sugere uma moderada multicolinearidade mas num nível menor do que o observado para a Alemanha.

d) Valores Reais vs. Preditos: O gráfico mostra uma boa correspondência entre os valores reais e os valores preditos pelo modelo, indicando que o modelo captura efetivamente a tendência dos dados.

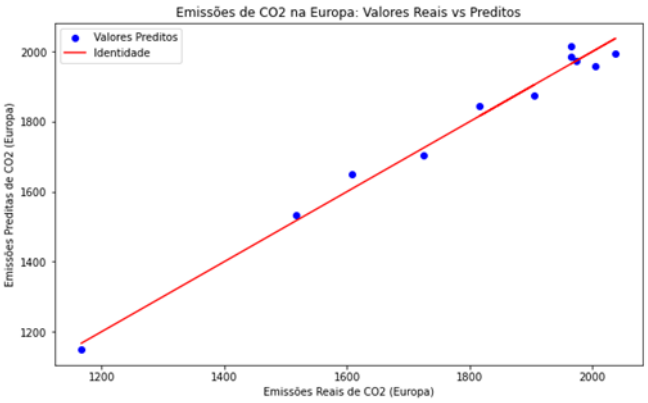

Figura 12 – Valores Reais vs Preditos

QQ Plot dos Resíduos: O QQ plot mostra que os resíduos seguem aproximadamente uma distribuição normal, apoiando a validade dos pressupostos do modelo de regressão.

Matriz de Correlação entre Variáveis Independentes: A matriz mostra que algumas correlações entre as variáveis são moderadas a fortes, o que sugere a presença de associações consideráveis entre as emissões de CO2 do carvão em diferentes países.

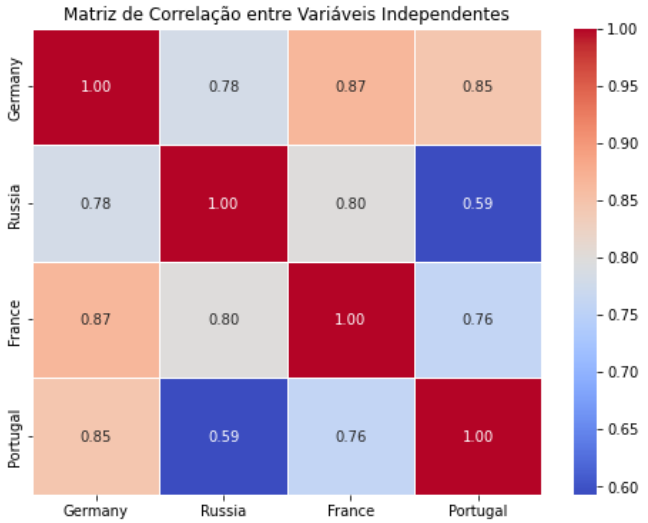

Figura 13 – Matriz de Correlação entre Variáveis Independentes

e) Realizamos uma estimativa das emissões de CO2 na Europa para o ano de 2015 com base nos dados disponíveis para os países selecionados. A comparação entre a estimativa e o valor real mostra uma boa concordância, sugerindo que o modelo é capaz de prever com precisão as emissões de CO2 na Europa.

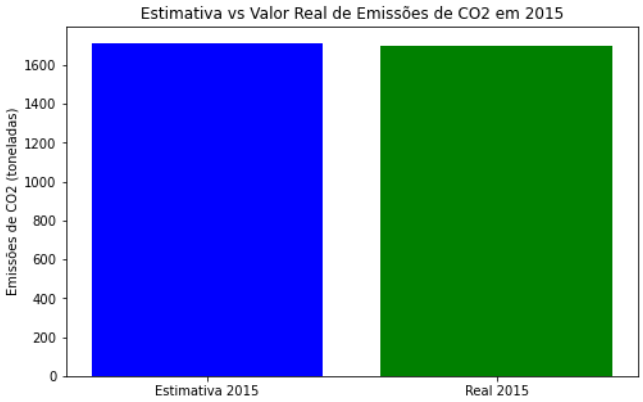

Figura 14 – Gráfico do exercício 4.3.2 alínea e)

IV. Conclusões

Neste estudo, exploramos uma variedade de técnicas estatísticas para analisar dados de emissões de CO2 em diferentes regiões do mundo.

Através de testes de hipóteses, identificamos padrões interessantes no comportamento das economias de Portugal e Hungria, embora não tenhamos encontrado diferenças significativas em relação ao Produto Interno Bruto (PIB).

Em seguida, ao investigar a correlação entre as emissões de CO2 do carvão em diferentes regiões, observamos relações complexas e diferenças entre essas variáveis.

Além disso, ao aplicarmos análises de regressão linear, buscamos entender como as emissões de CO2 em países específicos podem influenciar as emissões de CO2 em toda a região europeia.

Por fim, ao abordar o último exercício, construímos um modelo de regressão linear para estimar as emissões de CO2 na Europa com base nos dados de emissões de CO2 do carvão de países como Alemanha, Rússia, França e Portugal.

Este estudo fornece insights valiosos sobre as interações entre atividades económicas e emissões de CO2, destacando a importância da análise estatística na compreensão dos impactos ambientais globais. Estas descobertas podem ser úteis para formuladores de políticas e pesquisadores que trabalham para mitigar as mudanças climáticas e promover o desenvolvimento sustentável em todo o mundo.

References

[1] Wes Mckinney, Python for Data Analysis: Data Wrangling with pandas, NumPy,and Jupyter, 3rd edition, https://wesmckinney.com/book/ ,O'Reilly Media, 2022. (accessed April. 1, 2024)

[2] Douglas C. Montgomery, "Design and Analysis of Experiments, 8th edition. John Wiley & Sons," New York, 2013. https://mip.faperta.unri.ac.id/file/bahanajar/58219-2013-8ed-Montgomery-Design-and-Analysis-of-Experiments.pdf (accessed April. 5, 2024)

[3] João Matos (ISEP), "Regressão Linear" https://moodle.isep.ipp.pt/pluginfile.php/369619/mod_resource/content/1/ANADI-LEI-1718_RegressaoLinear.pdf (accessed April. 3 , 2024).

[4] Ana Madureira, João Matos, "Aulas T - Testes de Correlação" https://moodle.isep.ipp.pt/pluginfile.php/369616/mod_resource/content/1/Testes_de_Correlacao.pdf

[5] ChatGPT, 2023. https://chat.openai.com/ (accessed March. 26, 2024)

[6] Wikipedia contributors. "Shapiro–Wilk test." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 8 Feb. 2024. Web. 7 Apr. 2024.

[7] WES MCKINNEY, PYTHON FOR DATA ANALYSIS: DATA WRANGLING WITH PANDAS, NUMPY, AND JUPYTER, 3RD EDITION, HTTPS://WESMCKINNEY.COM/BOOK/ MEDIA, 2022

[8] Ana Madureira ," ANADI 1ª aula TP" https://moodle.isep.ipp.pt/pluginfile.php/364042/mod_resource/content/1/Aula_1_introduc%CC%A7a%CC%83o.pdf

[9] Ana Madureira, João Matos, "Testes de Hipóteses não Paramétricos" https://moodle.isep.ipp.pt/pluginfile.php/368358/mod_resource/content/1/TestesNaoParametricos.pdf

[10] JEM, AMD, TPA, "Testes de Hipóteses paramétricos TP2" https://moodle.isep.ipp.pt/pluginfile.php/366181/mod_resource/content/2/Aula2_TP2.pdf

[11] JEM, AMD, TPA, "Proposta de resolução da ficha TP4" https://moodle.isep.ipp.pt/pluginfile.php/371127/mod_resource/content/1/Proposta_de_resolucao_TP4.pdf