ANADI – Trabalho Prático 1

Gabriel Gonçalves
Departamento de Engenharia Informática
Instituto Superior de Engenharia do Porto
Porto, Portugal
1191296@isep.ipp.pt

Tiago Leite
Departamento de Engenharia Informática
Instituto Superior de Engenharia do Porto
Porto, Portugal
1191369@isep.ipp.pt

Francisco Bogalho
Departamento de Engenharia Informática
Instituto Superior de Engenharia do Porto
Porto, Portugal
1211304@isep.ipp.pt

***Abstract* — A Unidade Curricular de Análise de Dados em Informática (*ANADI*) propôs a resolução de um conjunto de exercícios, com o objetivo de permitir aos alunos aplicar e aprofundar os seus conhecimentos. Este artigo apresenta uma solução possível para os referidos exercícios, feita com a linguagem de programação *Python* no ambiente interativo *Jupyter* N*otebook*, incluindo as explicações dos processos adotados e a fundamentação teórica que suporta as decisões tomadas.**

Keywords — análise, dados, estatística, gráfico, testes de hipótese, inferência, correlação, regressão, amostra e variáveis.

I. Introdução

Com o grande aumento do volume de dados a serem processados, a área de Análise de Dados tem vindo a ganhar cada vez mais destaque. Neste contexto, os exercícios propostos para a Unidade Curricular de Análise de Dados em Informática (*ANADI*) surgem como uma oportunidade para explorar e aplicar conceitos fundamentais.

Este artigo começa com uma explicação teórica dos métodos utilizados nos exercícios, seguida pela descrição detalhada dos procedimentos adotados para resolver cada problema. Por fim, apresentaremos as conclusões gerais decorrentes deste trabalho analítico.

II. Introdução Teórica

A. Caraterísticas das Amostras

Para analisar os dados é necessário ter em conta determinados conceitos que ajudem a compreender e entender as conclusões tiradas.

1) Amostras Dependentes

Uma amostra dependente é aquela que é afetada pelos valores de outra amostra, ou seja, quando os valores de duas amostras se relacionam através de algum dado que partilhem.

2) Amostras Independentes

Uma amostra independente é aquela que não é afetada pelos valores de outra amostra, ou seja, quando duas amostras não têm qualquer tipo de relação entre as observações.

3) Distribuição Normal

A distribuição normal é um modelo que descreve a probabilidade de diferentes valores em um conjunto de dados, sendo contínua e sem espaços vazios. É determinada pela média e pelo desvio padrão, sendo esses dois parâmetros fundamentais para se entender como os dados estão distribuídos ao redor da média. Para se testar a normalidade de uma distribuição usa-se comumente o teste de *Shapiro-Wilk*.

4) Variância

A variância é uma medida estatística que indica o quão dispersos os valores de um conjunto de dados estão em relação à média.

5) Assimetria

A assimetria é uma medida estatística que representa as diferenças entre a variância á esquerda e á direita da média. Pode ser calculada com o coeficiente de assimetria de *Pearson* [1].

6) Autocorrelação

A autocorrelação é uma ferramenta matemática usada para encontrar padrões de repetição nos dados.

B. Testes

Há vários tipos de teste, sendo os analisados aqui, os testes de hipótese, que se dividem em testes paramétricos e testes não paramétricos, e os testes de correlação.

Os testes de hipótese são ferramentas estatísticas que permitem "investigar se uma determinada afirmação sobre um determinado parâmetro de uma população é verdadeiro ou falso." [2].

Estes testes avaliam uma hipótese, podendo se escolher a hipótese nula (*H0*) ou a alternativa (*H1*), dependendo do nível de significância (α), "dado um parâmetro desconhecido θ de uma população e um valor fixo β iremos considerar os seguintes três testes de hipótese" [2], e a partir deles escolher um:

- Teste bilateral: $H0$: $\theta = \beta$, então $H1$: $\theta \neq \beta$;
- Teste unilateral à esquerda: $H0$: $\theta \geq \beta$, então $H1$: $\theta < \beta$;
- Teste unilateral à direita: $H0$: $\theta \leq \beta$, o que implica que $H1$: $\theta > \beta$.

1) Testes Paramétricos

Os testes paramétricos são testes de hipótese que são realizados sobre amostras paramétricas, ou seja, quando os dados seguem uma distribuição normal, as variâncias são semelhantes, a autocorrelação é baixa e os dados são contínuos.

Alguns tipos de testes que podem ser aplicados no contexto de testes paramétricos são, por exemplo, teste *ANOVA* ou teste t [3].

2) Testes Não Paramétricos

Os testes não paramétricos são testes de hipótese que são aplicados quando as condições necessárias para se aplicar um teste paramétrico não são satisfeitas.

Alguns tipos de testes que podem ser aplicados no contexto de testes não paramétricos são, por exemplo, teste Binomial, teste *Wilcoxon* ou o teste *Kruskal-Wallis* [4].

3) Testes de Correlação

Os testes de correlação são usados para avaliar a relação entre duas vaiáveis. O teste avalia se existe associação e em casos positivo, retorna um valor de -1 a 1, que pode ser interpretado da seguinte forma [5]:

- Se estiver próximo de 1, as variáveis *X* e *Y* estão, positivamente, fortemente correlacionadas;
- Se r estiver próximo de −1, as variáveis *X* e *Y* estão, negativamente, fortemente correlacionadas;
- Se r estiver próximo de 0, então as variáveis *X* e *Y* estão fracamente correlacionadas.

Existem testes que podem ser usados para ambos os tipos de correlação, linear e ordinal. Para a linear pode-se usar o teste de correlação linear de *Pearson*. Já para a ordinal pode-se usar o teste de correlação ordinal de *Spearmen* ou o teste de correlação ordinal de *Kendall*.

C. Regressão Linear

A regressão linear é um método estatístico usado para modelar a relação entre uma variável dependente (ou resposta) e uma ou mais variáveis independentes (ou preditoras, ou explanatórias) [6]. A regressão linear é utilizada numa ampla variedade de áreas como na análise estatística, economia, engenharia, ciências sociais, biologia, entre outras. Algumas das aplicações mais comuns da regressão linear são, por exemplo, na análise de tendências, previsões, estudo de mercado e análise financeira [7].

1) Regressão Linear Múltipla

A regressão linear múltipla permite estudar a relação entre uma variável dependente (Y) e um conjunto de variáveis independentes (X_1, X_2, …, X_p, em que p > 1). A representação do modelo, é dada pela seguinte formula [6]:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \in$$

Onde [8]:

- Y é a variável dependente que estamos a tentar prever ou estimar;
- β_0 é o valor esperado de Y quando todas as variáveis independentes forem nulas;
- β_p é a variação esperada em Y dado um incremento unitário em X_p, mantendo-se constantes todas as demais variáveis independentes;
- ϵ é o erro aleatório não explicado pelo modelo.

Para que um Modelo de Regressão Linear Múltipla seja adequado, este deve satisfazer um conjunto de condições, sendo estas [6]:

- Normalidade – Os erros (resíduos), $\in$, têm uma distribuição normal $N(0, \sigma^2)$, ou seja, média igual a zero e variância constante;
- Homocedasticidade – A variância é constante para todos os níveis das variáveis independentes;
- Autocorrelação nula – Os erros são mutuamente independentes;
- Multicolinearidade – As variáveis $X_1, \ldots, X_p$ devem ser linearmente independentes.

O diagnóstico da Multicolinearidade pode ser feito utilizando vários métodos [6]:

- Através da análise dos valores próprios da matriz $C^T C$, onde C é a matriz das covariâncias. Sabemos que estamos perante multicolinearidade se os valores próprios forem pequenos.
- Utilizando o fator de inflação de variância (VIF), em que na prática, iremos considerar que há ausência de multicolinearidade quando VIF < 5.

D. Tipos de Testes

1) Teste de Shapiro-Wilk

O teste de *Shapiro-Wilk* é um teste de hipótese não paramétrico que é usado para verificar se uma amostra de dados segue uma distribuição normal [1].

Este teste é mais apropriado para amostras inferiores a 30, podendo ser aplicado a amostras superiores a esse número, a estatística teste é dada por [1]:

$$T(X) = \frac{\sum_{i=1}^{n} a_i X_{(i)}}{\sum_{i=1}^{n} \left(X_{(i)} - \overline{X}\right)^2}$$

Para se aplicar o teste em questão é necessário que os valores da amostra sejam independentes e contínuos.

2) Teste t de Student

O teste t de *Student* é um teste de hipótese paramétrico que é usado para comparar as médias de duas amostras independentes, a estatística teste é dada por [9]:

$$T(X) = \frac{\overline{X} - \mu_0}{\frac{S_X}{\sqrt{n}}} \sim T_{n-1}, \text{ com } S_X^2 = \frac{\sum_{i=1}^{n} (X_i - \mu_0)^2}{n-1}$$

Para se aplicar o teste em questão é necessário que a amostra seja paramétrica.

3) Teste de Wilcoxon

O teste de *Wilcoxon* é um teste de hipótese não paramétrico que é usado para comparar duas amostras pareadas. Usa-se quando as suposições necessárias para o teste t de *Student* não são atendidas.

Para se aplicar o teste de *Wilcoxon* começa-se por, para cada par de observações nas duas amostras, calcular a diferença entre essas observações. Essas diferenças podem ser positivas ou negativas. Em seguida, ordenamos essas diferenças em ordem crescente de magnitude e atribuímos *ranks* a elas, de 1 até o tamanho da amostra. Depois de termos os *ranks* para todas as diferenças, calculamos duas somas: uma soma dos *ranks* para as diferenças positivas e outra soma para as diferenças negativas. O valor final é o menor desses dois somatórios [1].

4) Teste de Levene

O teste de *Levene* é um teste de hipótese paramétrico que "é usado para verificar se k populações têm variâncias iguais" [2].

"O teste de *Levene* é definido por" [2]:

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2 \quad \text{versus} \quad \sigma_i^2 \neq \sigma_j^2 \quad \text{para algum par } (i,j)$$

"Dada uma v.a. *X* com tamanho *n* dividida em *k* subgrupos, com cada subgrupo com tamanho ni, a estatística teste é" [2]:

$$T(X) = \frac{n-k}{k-1} \frac{\sum_{i=1}^{k} n_i (Z_{i.} - Z_{..})^2}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (Z_{i,j} - Z_{i.})^2} \sim F_{k-1, N-k}$$

"onde Zi,j pode tomar uma das seguintes formas" [2]:

$Z_{i,j} = |X_{i,j} - \overline{X}_{i.}|$, onde $\overline{X}_{i.}$ é a média do *i*-ésimo grupo

$Z_{i,j} = |X_{i,j} - \widetilde{X}_{i.}|$, onde $\widetilde{X}_{i.}$ é a mediana do *i*-ésimo grupo

$Z_{i,j} = |X_{i,j} - X'_{i.}|$, onde $\widetilde{X}'_{i.}$ é a média truncada do *i*-ésimo grupo

e

$Z_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} Z_{i,j}$

$Z_{..} = \frac{1}{n}\sum_{i=1}^{k}\sum_{i=1}^{n_i} Z_{i,j}$

5) Teste de ANOVA

O teste *ANOVA* é um teste de hipótese paramétrico que é usado para comparar as médias de duas ou mais amostras independentes e verificar se existem diferenças estatísticas significativas [2].

O teste *ANOVA* é uma baseado no teste t de *Student* para amostras independentes e tem-se que a variável independente é categórica que define os grupos que serão comparados e que a variável dependente é uma variável numérica, cujas médias serão comparadas [2].

Para se puder aplicar o teste *ANOVA* é necessário existir alguns pressupostos [2]:

- A variável dependente deve ser contínua;
- A variável independente deve ter dois ou mais grupos independentes. Geralmente usa-se apenas um teste *ANOVA* para três ou mais grupos categóricos independentes uma vez que para dois grupos é mais cómodo fazer t-teste.
- As observações devem ser independentes;
- As observações não devem ter *outliers* significativos;
- A variável dependente deve ser normalmente distribuída para cada grupo;
- Deve existir homogeneidade de variâncias.

A estatística teste é dada por [2]:

$$T(X) = \frac{\frac{TSS}{k-1}}{\frac{RSS}{n-k}} \sim F_{k-1,n-k}$$

6) Teste de Kruskal-Wallis

O teste de *Kruskal-Wallis* é um teste de hipótese não paramétrico que é usado como alternativa ao teste ANOVA. "Deve-se utilizar quando a hipótese da normalidade for rejeitada ou se os tamanhos das amostras forem pequenas." [1]. "O objectivo do teste de *Kruskal-Wallis* é testar se uma dada variável qualitativa, designada de factor tem efeitos iguais sobre uma determinada variável quantitativa." [1].

A estatística teste é dada por [1]:

$$T = \frac{12}{n(n+1)}\sum_{i=1}^{k}\frac{T_i^2}{n_i} - 3(n+1) \sim \chi^2_{(k-1)}$$

7) Teste de Durbin-Watson

O teste de *Durbin-Watson* é um teste de hipótese paramétrico que é usado para verificar a presença de autocorrelação, é frequentemente utilizado em análises estatísticas, como regressão linear, simples ou múltipla.

Se o valor devolvido pelo teste *Durbin-Watson* for menor que 1.5, então existem sinais de autocorrelação positiva. Se o valor for maior que 2.5, então existem valores de autocorrelação negativa. Se o valor estiver contido entre 1.5 e 2.5, então não há autocorrelação [10].

A estatística teste é dada por [11]:

$$d = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}$$

8) Teste de Correlação Linear de Pearson

O teste de correlação linear de *Pearson* é um teste de correlação que é usado para medir o grau de relação linear entre duas variáveis [5].

Tem-se que [5]:

$$r(X,Y) = r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})\sum_{i=1}^{n}(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}}$$

A estatística de teste é dada por [5]:

$$T(X,Y) = r \times \sqrt{\frac{n-2}{1-r^2}} \sim T_{n-2}$$

A hipótese pode ser uma de três [5]:

- $H_0: r = 0 \; vs \; H_1: r = 0$
- $H_0: r = 0 \; vs \; H_1: r > 0$
- $H_0: r = 0 \; vs \; H_1: r < 0$

Existem alguns pressupostos para se poder aplicar este teste [5]:

- As variáveis devem ser contínuas e não devem existir *outliers* significativos;
- Deve existir uma relação linear entre as duas variáveis;
- As variáveis devem ter aproximadamente uma distribuição normal;
- E as variâncias devem ser iguais.

III. Métodos e Resultados Obtidos

Para a realização deste trabalho foi utilizado em todas as instâncias um nível de significância de 5% para todos os testes de hipótese efetuados.

A. Análise e exploração de dados

1) Exercício 1

No primeiro exercício, analisaram-se as emissões totais de CO_2 de Portugal entre 1900 e 2021. Inicialmente, os dados foram filtrados para incluir apenas as informações relevantes e limitar o período de interesse.

Em seguida, foi elaborado um gráfico de linha que mostra a evolução das emissões ao longo do tempo. O eixo *x* representa os anos, enquanto o eixo *y* reflete as emissões de CO_2 em milhões de toneladas.

Identificou-se que o pico de emissões ocorreu em 2005, para isso utilizou-se uma função do *pandas* que determinou o índice correspondente ao máximo valor de CO_2 e recuperou-se o ano associado a esse índice.

Esta análise observada na Figura 1 proporcionou uma compreensão detalhada das tendências das e missões de CO_2 de Portugal ao longo do tempo, bem como a identificação do ano de pico de emissões.

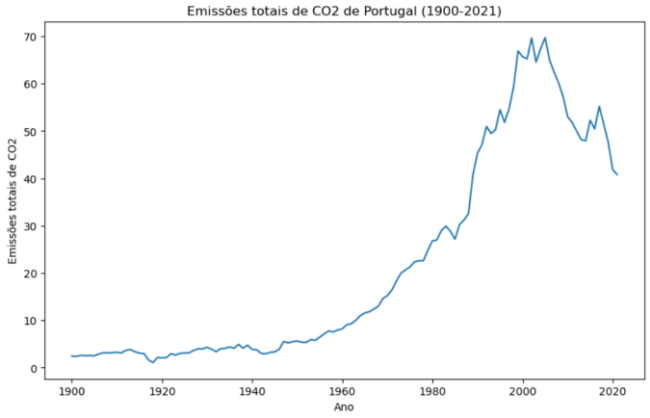

Figura 1 – Gráfico das emissões totais de CO_2 de Portugal (1900-2021)

2) Exercício 2

No segundo exercício comparam-se as emissões de CO_2 de Portugal, provenientes de: cimento, carvão, queima, gás, metano, óxido nitroso e do petróleo, entre 1900 e 2021.

Para isso inicialmente os dados foram filtrados para incluir apenas as informações relevantes e limitar o período de interesse.

Depois os dados, que se pretende comparar, foram adicionados ao gráfico com nomes indicativos do que representam, posteriormente legendou-se e mostrou-se o gráfico.

Ao analisar o gráfico presente na Figura 2, pode-se ver que a fonte mais consumida foi o petróleo e o seu pico de consumo foi em 2002. Também é visível que o metano e o óxido nitroso só começaram a ser utilizados em 1990 e pararam de o ser em 2019, similarmente o carvão desceu extremamente rápido encontrando-se num mínimo histórico em 2021.

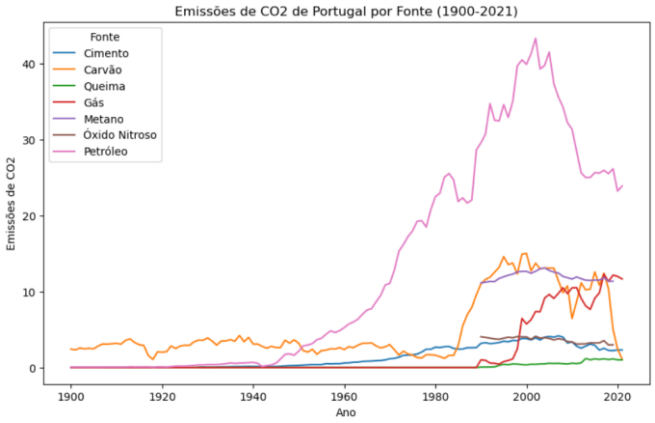

Figura 2 – Gráfico das emissões de CO_2 de Cimento, Carvão, Queima, Gás, Metano, Óxido Nitroso e Petróleo em Portugal, no período de 1900 a 2021.

3) Exercício 3

No terceiro exercício, construiu-se um gráfico que permite comparar as emissões de CO_2 *per capita* de Portugal com Espanha, no período de 1900 a 2021.

Primeiramente, filtrou-se os dados disponíveis para obtermos apenas os que são referentes a Portugal e Espanha, no período desejado. Através da análise dos resultados obtidos pela filtragem, não se identificou nenhum ano sem valores referentes às emissões de CO_2, nem sem a informação do número de habitantes, tanto no país de Portugal como no de Espanha. Desta forma, para a análise que se pretende realizar neste exercício, não houve necessidade de remover algum ano em que não houvesse a informação pretendida em ambos os países.

Em seguida, como a informação disponibilizada pelo ficheiro de dados era sobre emissões totais anuais de CO_2, medidas em milhões de toneladas, houve a necessidade de manipular estes dados para a representação pretendida. Dessa forma, alterou-se a representação de milhões de toneladas para simplesmente toneladas, através da multiplicação do valor de emissões de CO_2 de cada ano e país por um milhão (1 000 000). De seguida, para se obter os dados *per capita*, dividiu-se o novo valor obtido pela quantidade de habitantes que existiam no país nesse ano.

Por fim, construi-se um gráfico onde é possível visualizar a diferença das emissões de CO_2 *per capita* entre os países Portugal e Espanha.

Na Figura 3 é possível observar os resultados obtidos na análise realizada.

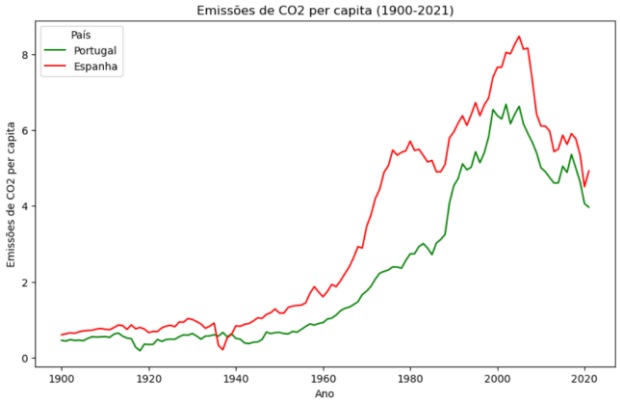

Figura 3 - Gráfico das emissões de CO_2 per capita nos países Portugal e Espanha, no período de 1900 a 2021.

Pela análise dos resultados obtidos, que podem ser visualizados na figura anterior, concluiu-se que no período de 1900 a 2021, Portugal na maioria dos anos desse período esteve sempre com um menor índice de emissões de CO_2 per capita em relação a Espanha. As únicas exceções que se identificou, foi no período de 1936 a 1939, em que Espanha conseguiu estar abaixo dos valores de Portugal.

4) Exercício 4

No quarto exercício, procedeu-se à análise comparativa das emissões de CO_2 provenientes do carvão entre os Estados Unidos, China, Índia, União Europeia (a 27) e Rússia, abrangendo o período de 2000 a 2021. Inicialmente, os dados foram filtrados para incluir apenas as informações relevantes desses países e limitar o intervalo temporal de interesse.

Os dados foram agrupados por ano e país, permitindo a criação de uma série temporal para cada país e a construção de um gráfico de linha para visualizar e comparar as emissões. Cada país foi representado por uma linha no gráfico, onde o eixo *x* representa o ano e o eixo *y* representa as emissões de CO_2 em milhões de toneladas originadas pelo carvão.

Esta análise observada na Figura 4 proporcionou *insights* sobre as tendências das emissões de CO_2 do carvão entre os países, contribuindo para uma compreensão mais ampla do seu papel nas emissões globais de CO_2.

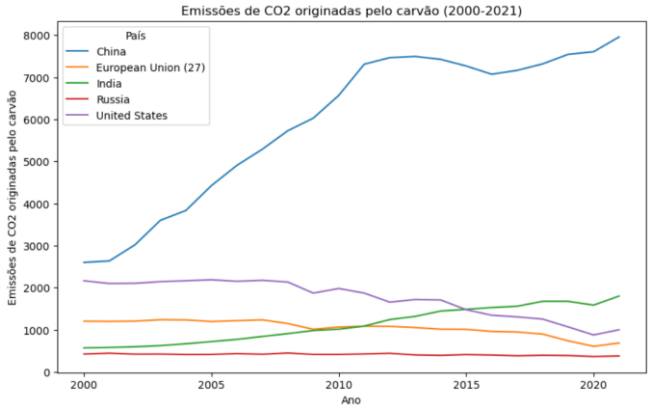

Figura 4 – Gráfico das emissões de CO_2 originadas pelo carvão (2000-2021)

Analisando o gráfico percebe-se que existe um aumento significativo nas emissões de CO_2 da China, enquanto os outros países mantêm-se estáveis ou diminuem ligeiramente

de 2000 a 2021. Este padrão destaca o papel crescente da China nas emissões globais de CO_2 originadas pelo carvão.

5) Exercício 5

No quinto exercício construiu-se uma tabela que indica, para cada uma das regiões: Estados Unidos, China, Índia, União Europeia (a 27) e a Rússia, as médias das emissões de CO_2 devidas a cimento, carvão, queima, gás, metano, óxido nitroso e do petróleo no período 2000-2021.

Para isso inicialmente os dados foram filtrados para incluir apenas as informações relevantes e limitar o período de interesse.

Depois agruparam-se os dados em análise, ou seja, cimento, carvão, queima, gás, metano, óxido nitroso e petróleo, calculou-se a média, com três casas decimais, para cada um deles e reformatou-se a tabela para melhor visualização.

Na Figura 5 está representada a tabela construída com a média de cada fonte em cada país e nela pode-se observar que a China tem valores substancialmente mais altos que os restantes á parte de queima (*flaring*) e gás, sendo que as emissões da queima são praticamente nulas. A Rússia é o país que tem os valores mais baixos.

	country	cement_co2	coal_co2	flaring_co2	gas_co2	methane	nitrous_oxide	oil_co2
0	China	599.141	5920.797	1.722	287.021	1015.726	476.530	1116.257
1	European Union (27)	81.488	1049.236	21.132	774.871	407.444	238.482	1374.161
2	India	91.512	1123.795	2.661	92.464	617.360	228.242	469.662
3	Russia	21.837	413.504	43.061	766.698	599.007	58.484	353.289
4	United States	40.055	1750.037	52.728	1364.198	639.154	259.030	2379.692

Figura 5 - Tabela que indique, para cada uma das regiões: Estados Unidos, China, Índia, União Europeia e a Rússia, as médias das emissões de CO_2 devidas a cimento, carvão, queima, gás, metano, óxido nitroso e do petróleo no período 2000-2021.

B. Inferência Estatística

1) Exercício 1

No primeiro exercício desta secção, testou-se se a média do produto interno bruto de Portugal foi superior à média do produto interno bruto da Hungria no período de 1900 a 2021. Identificou-se que neste exercício estávamos perante amostras emparelhadas, pois os dados dos diferentes anos são dependentes devido ao produto interno bruto ser analisado nos mesmos dois países nos diferentes anos.

Primeiramente, começou-se por escolher uma amostra aleatória de anos no período de 1900 a 2021, em que esta tem um tamanho igual a 30.

De seguida, filtrou-se os dados para termos apenas os que são referentes a Portugal e à Hungria. Também se filtrou os dados para obtermos aqueles onde o valor do produto interno bruto não é nulo em ambos os países para o mesmo ano, e onde os anos que apenas são tomados em consideração fazem parte da amostra aleatória. Desta forma, eliminamos os registos dos anos em que não existem dados em ambos os países. Com esta análise, identificou-se que passamos de uma amostra de tamanho 30, para uma amostra com apenas 26 de tamanho. Isto deve-se a que na amostra aleatória inicial de tamanho 30, haviam 4 valores referentes a anos em que não haviam dados para o produto interno bruto nos dois países em simultâneo.

Posteriormente, efetuou-se um estudo dos dados agora existentes, para percebermos se era possível resolver este problema com a utilização de um teste paramétrico (teste-t, com amostras emparelhadas). Para tal, fez-se um estudo à normalidade dos dados através da realização de um teste de *Shapiro*. Tem-se as hipóteses:

H_0: Os dados seguem uma distribuição normal

H_1: Os dados não seguem uma distribuição normal

O resultado obtido deste teste foi de 0.00017, pelo que com um nível de significância de 5%, rejeita-se a hipótese H_0, pelo que não se deve utilizar o *teste-t*, mas sim usar antes o teste de *Wilcoxon*.

Contudo, para aplicar o teste de *Wilcoxon*, foi necessário realizar o teste à simetria dos dados de forma a verificar se o teste deve ser efetuado à média ou apenas à mediana. Para isso, calculou-se o coeficiente de assimetria (*skewness*). O valor que se obtém neste cálculo, irá dar a seguinte informação [1]:

- $|skewness| < 0.1$ – Distribuição simétrica;
- $0.1 < |skewness| < 1$ – Distribuição moderadamente assimétrica;
- $|skewness| > 1$ – Distribuição fortemente assimétrica.

O valor absoluto do coeficiente de assimetria obtido foi de 0.60285, pelo que é inferior a 1, mas superior a 0.1, sendo que então consideramos os dados moderadamente assimétricos. Consequentemente, consideramos que a mediana é muito semelhante à média e o teste de *Wilcoxon* pode ser feito á média.

Para realizar o teste de *Wilcoxon*, considerou-se as seguintes hipóteses:

$H_0: \eta_{Portugal} = \eta_{Hungria} vs\ H_1: \eta_{Portugal} > \eta_{Hungria}$

O valor de prova obtido foi de 0.07136, pelo que com um nível de significância de 5%, não há evidência estatística de que a média do produto interno bruto de Portugal foi superior à da Hungria, no período de 1900 a 2021. Desta forma, não se rejeita a hipótese nula (H_0).

2) Exercício 2

No segundo exercício, testa-se se a média do produto interno bruto (*gdp*) de Portugal foi superior à média do produto interno bruto da Hungria no período 1900-2021, para isso vamos ter de aplicar um teste t de *Student* independente, para se decidir se se vai rejeitar a hipótese nula, sendo que:

$H_0: \eta_{Portugal} > \eta_{Hungria} vs\ H_1: \eta_{Portugal} <= \eta_{Hungria}$

Começou-se por se escolher duas amostras aleatórias de anos, foram escolhidos 12 anos do intervalo de 1900 a 2021, ambas as amostras forma geradas com *seed's* diferentes.

Depois os dados foram filtrados para incluir apenas as informações relevantes e limitar o período de interesse.

Para se aplicar o teste t de *Student* é preciso que as amostras tenham uma distribuição normal e a mesma variância, para se descobrir isso é preciso aplicar um teste de *Shapiro-Wilk* a cada amostra e um teste de *Levene*.

Com todos os dados prontos a serem analisados é executado um teste de *Shapiro-Wilk* para cada amostra, de forma a descobrir se se rejeita a hipótese nula:

H_0: Os dados seguem uma distribuição normal

H_1: Os dados não seguem uma distribuição normal

O *p-value* de Portugal foi de 0,115 e o da Hungria foi de 0,188, como ambos os valores são maiores que o nível de significância, de 0,05, não se rejeita a hipótese nula em ambos os casos, ou seja, ambas as amostras seguem uma distribuição normal.

Posteriormente foi aplicado o teste de *Levene* a cada uma das amostras para se descobrir se se rejeita a hipótese nula:

H_0: As amostras têm variâncias iguais

H_1: As amostras não têm variâncias iguais

Como o *p-value* foi de 0,567, superior ao nível de significância, não se rejeita a hipótese nula, ou seja, as variâncias são iguais.

Sabendo que ambas as amostras são normais e têm a mesma variância pode-se aplicar o teste t de *Student*, desta forma obtém-se o seu *p-value* de 0,532, como o valor é maior que o nível de significância não se rejeita a hipótese nula, logo não podemos concluir que a média do GDP de Portugal é superior á média do GDP da Hungria.

3) Exercício 3

No terceiro exercício, foram realizados testes estatísticos para investigar possíveis diferenças nas emissões totais de CO_2 entre as regiões dos Estados Unidos, Rússia, China, Índia e União Europeia (a 27).

Inicialmente, o teste de *Shapiro-Wilk* foi utilizado para verificar a normalidade das distribuições de CO_2 em cada região, revelando que apenas as emissões da Rússia seguiram uma distribuição normal.

Dado que nem todas as distribuições foram consideradas normais, optou-se por realizar o teste *Kruskal-Wallis*, uma alternativa não paramétrica ao teste *ANOVA*, para avaliar as diferenças nas médias das emissões de CO_2 entre as regiões. O resultado do teste indicou que existem diferenças significativas nas emissões de *CO2* entre as regiões consideradas.

Posteriormente, para identificar onde ocorreram estas diferenças, realizou-se o teste *post-hoc* de *Dunn*. A tabela resultante observada na Figura 6 forneceu os p-valores ajustados para cada par de regiões, onde um valor menor que 0,05 indica uma diferença significativa. Por exemplo, observou-se diferenças significativas nas emissões de CO_2 entre a Índia e a União Europeia (a 27), entre a Índia e os Estados Unidos, e entre a Rússia e os Estados Unidos.

Resultado do teste post-hoc de Dunn:

	China	European Union (27)	India	Russia	United States
China	1.000000	0.685863	0.397824	0.685863	0.102012
European Union (27)	0.685863	1.000000	0.016236	0.340554	0.685863
India	0.397824	0.016236	1.000000	0.685863	0.000272
Russia	0.685863	0.340554	0.685863	1.000000	0.023025
United States	0.102012	0.685863	0.000272	0.023025	1.000000

Figura 6 – Resultado do teste post-hoc de Dunn

Estes procedimentos permitiram uma análise detalhada das disparidades nas emissões de CO_2 entre as regiões selecionadas.

C. Correlação e Regressão

1) Exercício 1

No primeiro exercício desta secção, realizou-se a análise das emissões de CO_2 provenientes do carvão entre o ano de 2000 e o de 2021, abrangendo diversas regiões, sendo elas África, Ásia, América do Sul, América do Norte, Europa e Oceânia. Inicialmente, filtrou-se os dados para incluir apenas as informações relativas a essas regiões e também para limitar o intervalo temporal de interesse. De seguida, agrupou-se os dados por ano e país, combinando as emissões de CO_2 originadas pelo carvão, permitindo a construção de uma série temporal para cada região, evidenciando a evolução das emissões ao longo do tempo.

Posteriormente, construiu-se uma tabela de correlação para investigar as relações entre as emissões de CO_2 provenientes do carvão nas diferentes regiões. Para isso, utilizou-se o coeficiente de correlação de *Pearson*. Na Figura 7 é possível observar os resultados obtidos.

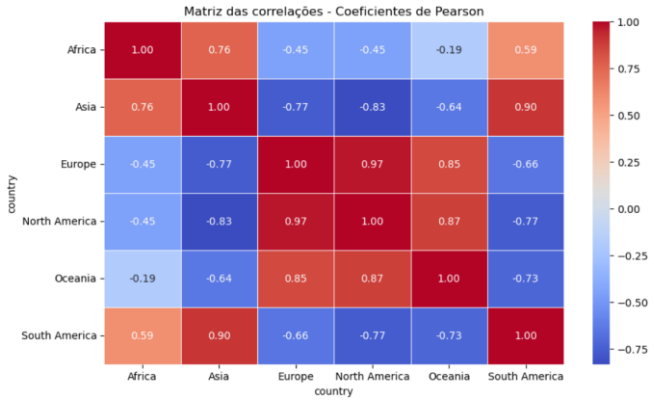

Figura 7 – Matriz dos coeficientes de correlação de Pearson entre os diferentes continentes

Os resultados obtidos na matriz de correlação fornecem perceções úteis sobre a relação entre as emissões de CO_2 provenientes do carvão nas diferentes regiões. Essas informações são essenciais para compreender os padrões globais de emissões de CO_2 e identificar possíveis tendências e padrões de comportamento ao longo do tempo.

2) Exercício 2

As próximas alíneas terão por base as seguintes variáveis independentes *X1*, *X2*, *X3* e *X4* sendo que representam as emissões de CO_2 provenientes do carvão nos anos pares do século *XXI* na Alemanha, Rússia, França e Portugal, respetivamente. E variável dependente *Y* que representa a emissão de CO_2 provenientes do carvão nos anos pares do século *XXI* na Europa.

a) Alínea a)

Para a alínea *a)* do segundo exercício, pretende-se encontrar o modelo de regressão linear. Para isso, concatena-se todas as variáveis independentes num único *DataFrame*, depois renomeia-se as linhas do *DataFrame* pela mesma ordem que as variáveis independentes foram concatenadas, ou seja, *Germany*, *Russia*, *France* e *Portugal*. Posteriormente adiciona-se uma linha de constante para o termo de intercetação e ajusta-se o modelo de regressão linear usando *Ordinary Least Squares* (OLS), obtendo assim a tabela presente na Figura 8.

OLS Regression Results

Dep. Variable:	coal_co2	R-squared:	0.984
Model:	OLS	Adj. R-squared:	0.974
Method:	Least Squares	F-statistic:	93.23
Date:	Wed, 03 Apr 2024	Prob (F-statistic):	1.57e-05
Time:	03:03:47	Log-Likelihood:	-53.778
No. Observations:	11	AIC:	117.6
Df Residuals:	6	BIC:	119.5
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-417.1743	328.717	-1.269	0.251	-1221.516	387.168
Germany	2.9544	0.699	4.229	0.006	1.245	4.664
Russia	2.2726	1.021	2.226	0.068	-0.226	4.771
France	6.9451	2.336	2.973	0.025	1.228	12.662
Portugal	-6.3903	7.537	-0.848	0.429	-24.832	12.051

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

Figura 8 - Modelo de Regressão Linear

b) Alínea b)

Para a alínea *b)* do segundo exercício, realizou-se uma análise abrangente para avaliar a adequação do modelo de regressão linear aos dados disponíveis. Inicialmente, calculou-se os resíduos do modelo, que representam a diferença entre os valores observados e os previstos e são essenciais para verificar o ajuste do modelo aos dados.

Para avaliar a normalidade dos resíduos, foram utilizados dois métodos: o gráfico *Q-Q* apresentado na Figura 9 e o teste de *Shapiro-Wilk*. Este teste indicou que os resíduos seguem uma distribuição normal, pois o *valor-p* associado foi superior a 0.05.

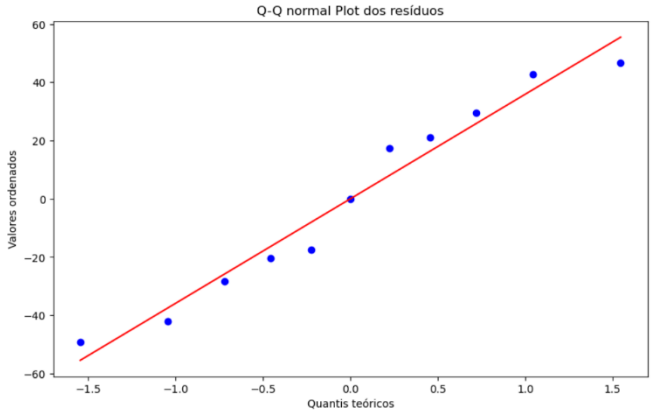

Figura 9 - Gráfico Q-Q de Normalidade dos Resíduos

Além disso, foi verificada a homocedasticidade dos resíduos por meio de um gráfico representado na Figura 10 de resíduos *vs* os valores previstos pelo modelo. No entanto, não se observou a homocedasticidade, pois não houve simetria em relação à reta y = 0, indicando uma variância não constante dos resíduos.

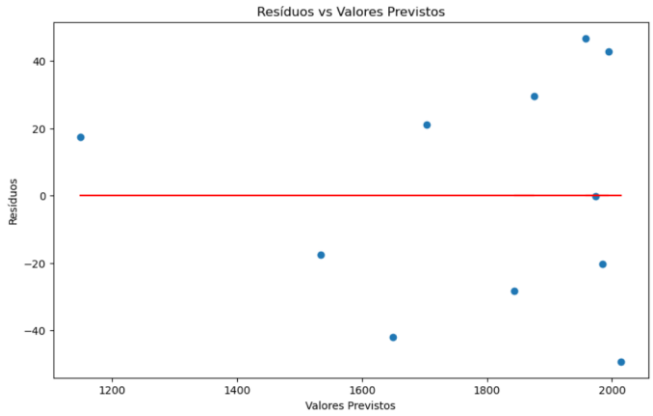

Figura 10 - Gráfico de Resíduos vs Valores Previstos

Por fim, o teste de *Durbin-Watson* foi realizado para verificar a presença de autocorrelação nos resíduos. Este teste indicou que não há autocorrelação nos resíduos, pois o valor obtido foi superior a 1.5 e inferior a 2.5.

Os resultados destas análises indicam que o modelo de regressão linear se ajusta bem aos dados, uma vez que os resíduos apresentam distribuição normal e não há autocorrelação. No entanto, a não verificação da homocedasticidade dos resíduos sugere a necessidade de uma investigação mais aprofundada para melhorar o modelo.

c) Alínea c)

Para a alínea c) do segundo exercício, realizou-se uma análise para verificar a existência de colinearidade entre as variáveis independentes utilizadas no modelo de regressão linear.

Utilizou-se o método do fator de inflação de variância (*VIF*) para calcular a colinearidade entre as variáveis independentes. Na Figura 11 é possível visualizar os resultados obtidos.

	variaveis	VIF
0	Germany	328.558342
1	Russia	138.679515
2	France	56.656309
3	Portugal	40.490640

Figura 11 - Fator de inflação de variância obtido para as diferentes variáveis

Através da análise dos resultados obtidos, concluiu-se que estes revelam que todas as variáveis independentes apresentaram valores de VIF significativamente elevados, indicando uma forte multicolinearidade entre elas.

Esses resultados sugerem que as variáveis independentes estão altamente correlacionadas entre si, o que pode afetar a precisão e interpretabilidade do modelo de regressão linear.

Nesse sentido, a presença de multicolinearidade deve ser considerada ao interpretar os coeficientes estimados do modelo e ao fazer previsões com base nele. Medidas para lidar com a multicolinearidade, como a exclusão de variáveis altamente correlacionadas ou o uso de métodos de regressão que reduzem o número de variáveis a um conjunto menor de componentes não correlacionados, podem ser utilizadas para melhorar a qualidade do modelo [12].

d) Alínea d)

Para a alínea *d)* do segundo exercício, pretende-se comentar o modelo obtido tendo em conta todas as características relevantes para a qualidade do modelo.

O *R-squared* quantifica o quanto as variações nas variáveis independentes estão associadas às variações na variável dependente. Neste caso, o *R-squared* é alto (0.984), o que sugere que o modelo explica 98.4% da variabilidade na variável dependente (coal_co2). Isso indica que o modelo tem um bom ajuste aos dados.

O adj. *R-squared* é uma versão corrigida do *R-squared* que leva em consideração o número de variáveis independentes no modelo de regressão. Neste caso, o valor é 0.974, o que ainda é alto e indica que o modelo provavelmente não está superestimando devido à inclusão de muitos preditores.

O *R-squared* ser mais alto que o adj. *R-squared* significa que o modelo inclui variáveis independentes adicionais que não contribuem significativamente para explicar a variabilidade na variável dependente.

O *F-statistic* avalia se pelo menos uma das variáveis independentes incluídas no modelo tem um efeito significativo na variável dependente. O valor alto do *F-statistic* (93.23) e a baixa probabilidade (1.57e-05) associada sugerem que o modelo como um todo é significativo.

Os valores na coluna *coef* significam um aumento de unidades em "coal_co2" para cada unidade de aumento.

A coluna P>|t| diz-nos se um coeficiente é estatisticamente significativo, ou seja, se P>|t| é menor que alfa (0.05).

O *Omnibus* é um teste de normalidade geral, enquanto o *Prob* (*Omnibus*) é um teste de normalidade específico. O *Omnibus* é um teste de qui-quadrado que compara a distribuição dos valores de resíduos com uma distribuição normal. O *Prob* (*Omnibus*) dá a probabilidade de que a distribuição dos resíduos seja normal. Se o valor de *Prob* (*Omnibus*) for menor que 0,05, a distribuição dos resíduos não é normal. Se o valor de *Prob* (*Omnibus*) for maior que 0,05, a distribuição dos resíduos é normal. Como *Prob* (*Omnibus*) é de 0,359, a distribuição dos resíduos é normal.

O *Skew* mede a assimetria dos resíduos. Um valor perto de zero como -0.026 significa que os resíduos têm uma distribuição simétrica.

O *Kurtosis* mede a forma da distribuição dos resíduos. Um valor de 1.631 significa que a distribuição dos resíduos é um pouco mais acentuada nas caudas.

O *Durbin-Watson* mede a autocorrelação dos resíduos, o valor de 1.577 sugere que há pouca autocorrelação nos resíduos.

O *Jarque-Bera* (*JB*) mede normalidade dos resíduos. Como o *Jarque-Bera* (*JB*) (0.860) é maior que o *Prob* (*JB*) (0.650), a distribuição dos resíduos é normal.

O Cond. no. representa a multicolinearidade entre as variáveis independentes no modelo. Como o número é maior que 30 (1.35e4), há uma alta multicolinearidade.

Com base na análise dos valores do modelo de regressão linear, podemos concluir que o modelo apresenta um alto ajuste aos dados, conforme evidenciado pelo alto *R-squared* (0.984) e adj. *R-squared* (0.974). Isso sugere que o modelo explica uma percentagem significativa da variabilidade na variável dependente (coal_co2) e não está superestimado devido à inclusão de muitos preditores. O valor elevado do *F-statistic* (93.23) e a baixa probabilidade associada indicam que o modelo como um todo é estatisticamente significativo. Além disso, os testes de normalidade dos resíduos (*Omnibus*, *Skew*, *Kurtosis*, *Durbin-Watson* e *Jarque-Bera*) sugerem que os pressupostos do modelo são atendidos, com distribuição normal dos resíduos e pouca autocorrelação. No entanto, deve-se ter cautela devido à alta multicolinearidade entre as variáveis independentes.

e) Alínea e)

No âmbito da alínea *e)*, estimou-se a emissão de CO_2 proveniente do carvão na Europa em 2015. Para tal, utilizou-se as variáveis independentes *X1*, *X2*, *X3* e *X4*, que representam as emissões nos anos pares do século *XXI* na Alemanha, Rússia, França e Portugal, respetivamente.

Os valores destas variáveis independentes para 2015 foram obtidos através dos dados disponíveis e inseridos num modelo de regressão linear ajustado previamente, incluindo uma constante para representar o termo de intercepção do modelo.

O modelo previu a emissão de CO_2 na Europa em 2015 como aproximadamente 1713.04 milhões de toneladas, com um erro absoluto de cerca de 10.80 milhões de toneladas em comparação com o valor real de aproximadamente 1702.24 milhões de toneladas.

Estes resultados indicam uma previsão precisa do modelo de regressão linear para as emissões de CO_2 na Europa em 2015.

IV. Conclusão

Durante a resolução dos exercícios, foi possível aprimorar os conhecimentos sobre Análise de Dados, com especial ênfase em temas como Inferência Estatística, Correlação e Regressão Linear.

Além disso, adquiriu-se conhecimentos sobre a importância de um bom tratamento inicial dos dados e a identificação das suas características, permitindo uma análise mais precisa do conjunto de dados. Notou-se também um avanço no uso de gráficos para a análise de dados que permitem extrair conclusões sobre a dispersão dos dados e a identificação de possíveis *outliers*.

Em suma, a experiência adquirida através da resolução dos exercícios foi extremamente valiosa. Não só permitiu a aplicação prática de conceitos teóricos, como também destacou a importância da análise de dados no mundo atual. A capacidade de interpretar e analisar dados é uma habilidade essencial em muitos campos e indústrias, portanto, aprofundar estes conhecimentos e habilidades é de grande relevância para qualquer profissional.

Referências

[1] A. Madureira e J. Matos, «Aulas Teóricas-Testes de Hipóteses Não Paramétricos».

[2] A. Madureira e J. Matos, «Aulas Teóricas-Testes de Hipóteses Paramétricos».

[3] Paula Villasante, «Testes paramétricos: definição e características». Acedido: 5 de Abril de 2024. [Em linha]. Disponível em: https://amenteemaravilhosa.com.br/testes-parametricos/

[4] Paula Villasante, «Testes não paramétricos: definições e tipos». Acedido: 5 de Abril de 2024. [Em linha]. Disponível em: https://amenteemaravilhosa.com.br/testes-nao-parametricos/

[5] A. Madureira e J. Matos, «Aulas T-Testes de Correlação».

[6] J. Matos, «Regressão Linear - Análise de Dados em Informática». 2024.

[7] «O que é regressão linear? -Maturidade». Acedido: 6 de Abril de 2024. [Em linha]. Disponível em: https://mathority.org/pt/regressao-linear/

[8] • Definição, «Regressão Linear Múltipla Econometria Alexandre Gori Maia Ementa».

[9] A. Madureira e J. Matos, «Aulas Teóricas-Testes de Hipóteses Paramétricos».

[10] «ANADI-LEI-1718_RegressaoLinear».

[11] «Durbin–Watson statistic - Wikipédia». Acedido: 6 de Abril de 2024. [Em linha]. Disponível em: https://en.wikipedia.org/wiki/Durbin%E2%80%93Watson_statistic

[12] «Basta! Lidando com a multicolinearidade na análise de regressão». Acedido: 7 de Abril de 2024. [Em linha]. Disponível em: https://blog.minitab.com/pt/basta-lidando-com-a-multicolinearidade-na-analise-de-regressao