Coimbra
Business School
Politécnico de Coimbra

FABÍOLA DE SOUZA QUEIROZ

MARIA INÊS ALMEIDA

OLEKSANDRA KUKHARSKA

PROPOSTA TEMA: EVOLUÇÃO DOS PREÇOS DO GÁS NATURAL EM PORTUGAL COMPARADO AOS PAÍSES DA UE

Descrição geral

Para o projecto DCBD, usaremos dados abertos da Eurostat (Gabinete de Estatísticas da União Europeia), onde incluiremos a etapa de Data Mining para analisar o Tarifário do Setor do Gás Natural em Portugal e traçar uma previsão. O objetivo é fomentar a existência de padrões utilizando agentes de mercado (variáveis explicativas) e comparar ao cenário atual dos demais países da UE considerando a situação geopolítica existente.

Unidade de Análise

Moeda: Euro (€) País: Portugal

Variáveis explicativas

- Inflação ou **deflator do PIB**: Instrumento utilizado para medir a inflação registrada num determinado espaço econômico. Trata-se de um indicador de periodicidade anual que integra os preços de todos os bens e serviços que existem numa economia;
- Custos de aquisição do gás natural;
- Procura do gás (consumo do GN em ano civil);
- Regulamento Tarifário GN, taxas de juros (Euribor, Spread)

Tabelas Sobre política da EU : Balança comercial líquida de produtos energéticos - % do PIB (tipsen10)
Estatísticas de Energia : Preços do gás para consumidores domésticos e não domésticos.

Variável de Saída

Evolução dos preços do gás natural para consumidores domésticos e não domésticos em Portugal em relação aos demais países União Europeia (UE) considerando a situação geopolítica atual.

Descrição simplificada

- Atividade preditiva com análise de tendências;

Unidade Curricular de Data Mining

Mestrado em Análise de Dados e Sistemas de Apoio à Decisão

Ano Letivo 2021 - 2022

Evolução dos preços do gás natural em Portugal comparando com os países da UE

Autor(es)

Fabíola de Souza Queiroz

Maria Inês Vicente

Oleksandra Kukharska

Elaborado em

19/07/2022

Coimbra
Business School
Politécnico de Coimbra

Índice

Coimbra
Business School
Politécnico de Coimbra

Índice de figuras

Lista de siglas e acrónimos

DCBD - Descoberta de Conhecimento em base de dados

CRISP-DM - CRoss-Industry Standard Process for Data Mining

UE - União Europeia

DGEG - Direção Geral de Energia e Geologia

KDD - Knowledge Discovery in Databases

CDD - índice do grau de arrefecimento diário

HDD - índice de graus-dia de aquecimento

Coimbra
Business School
Politécnico de Coimbra

Introdução

Neste projecto foi utilizado o processo de DCBD, com a mineração dos dados abertos da Eurostat (Gabinete de Estatísticas da União Europeia), seguindo a metodologia de data mining CRISP-DM e outras técnicas para a explicação do tarifário do Setor do Gás Natural em Portugal e traçar uma previsão. O objetivo é fomentar a existência de padrões utilizando agentes de mercado (variáveis explicativas) e comparar ao cenário atual dos demais países da UE considerando a situação geopolítica existente. De acordo com a ilustração abaixo, escalamos o tema analisado e de acordo com as etapas adequamos as pesquisas ao fluxo proposto (Chapman et al, 2000), adaptado por Dataprev:

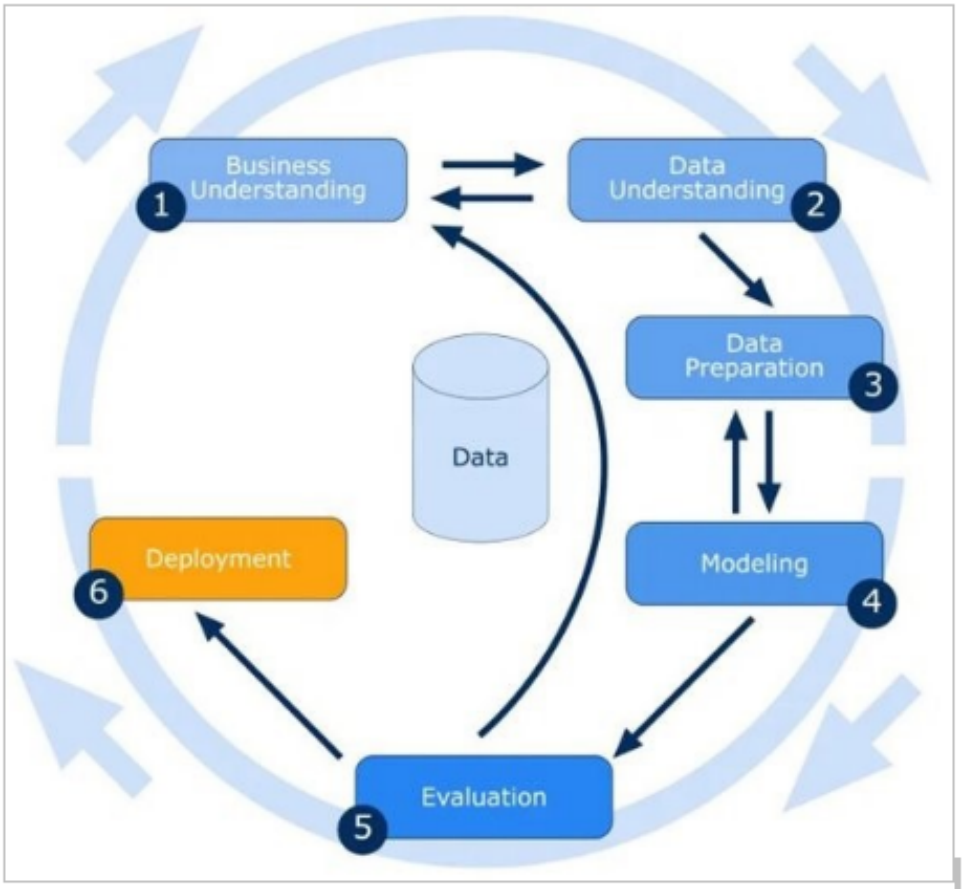

Figura 1: Etapas da metodologia CRISP - DM - The CRISP-DM life cycle, (Chapman et al, 2000) - Adaptado por Dataprev

Etapa 1- ***Entendimento do Negócio (Business Understanding):*** O problema de negócio em tese visa analisar o aumento dos preços e do consumo do gás natural, fazendo uma comparação a nível da União Européia, levando em conta o cenário da atual situação Geopolítica na Europa da Guerra Ucrânia X Rússia.

Etapa 2 - *Entendimento dos Dados (Data Understanding):* Existem vários fatores que impactam diretamente nos preços do gás natural, mas, após um estudo do objetivo geral constatou-se características significativas que alteram o tarifário: Graus-dias de aquecimento e resfriamento, importação e exportação, banda de consumo e a evolução histórica dos preços.

Etapa 3 - *Preparação dos Dados (Data Preparation):* Para preparar os dados para a modelagem avaliamos todas as bases de dados abertos sobre o tema e foi possível obter os dados brutos iniciais de uma única fonte(Eurostat), separados por variável, que passaram pela limpeza e transformação na próxima etapa;

Etapa 4 - *Modelagem (Modeling):* Aplicar técnicas de modelagem com o problema a ser resolvido (Próxima etapa do projeto).

Etapa 5 - *Avaliação (Evaluation):* Realização de testes com o modelo gerado para validar se atendem às necessidades do negócio(Próxima etapa do projeto).

Etapa 6 - *Utilização ou Aplicação (Deployment):* Apresentação dos resultados da modelagem para a tomada de decisão (Próxima etapa do projeto).

1- Entendimento do Negócio

1.1 Determinar objetivos do Negócio

A versatilidade do gás natural estimula e facilita a economia por isso a procura cada vez mais elevada a nível mundial, assim como o interesse público na regulação desses mercados extremamente elevados.
Este projeto tem o intuito de demonstrar como a regulação do preço permite avaliar a evolução do setor energético, considerando determinados cenários com dados estatísticos da Eurostat a nível da UE sobre decisões e planejamentos para avaliar o desenvolvimento em termos de energia, além destacar a harmonização das estatísticas em estreita cooperação com as autoridades nacionais.
O objetivo será, então, analisar a evolução dos preços com base em variáveis que descrevem as mudanças de comportamento de consumo, importação e exportação válidas para este problema.

1.1.1 Variável de Saída

Evolução dos preços do gás natural para consumidores domésticos e não domésticos em Portugal em relação aos demais países União Europeia (UE) considerando a situação geopolítica atual.

1.2 Funcionamento do Mercado de gás Natural

O setor de gás natural dos países Europeus, delineado por consumidores e fornecedores, deve estabelecer entidades reguladoras que administram a exploração, transporte, instalações de armazenamento, terminais e a distribuição de gás. Sabemos que Portugal não possui reservas de gás natural e importa a totalidade do que consome e possui a menor capacidade de armazenamento. Um mercado de gás padrão onde há produção , consumo e fornecimento segue o seguinte processo de regulação segundo a ERSE:

- Produção: Extração do subsolo através de poços;
- Transporte e armazenamento: Transportado em embarcações e encaminhado para pontos de armazenamento.
- Distribuição e consumo: As redes de distribuição asseguram o abastecimento direto dos consumidores domésticos e empresas, e por meio de contadores de gás permitem informar ao consumidor a energia consumida para facturação.

Figura 2: Funcionamento do gás, desde o armazenamento até ao consumo. ERSE - Entidade Reguladora dos Serviços Energéticos - " GÁS: COMO FUNCIONA?"

As negociações em Portugal são feitas, em sua maioria, através de cláusulas *take-or-pay,* onde determinam-se contratualmente quantidades mínimas de gás natural que satisfaçam a viabilidade económica e incentivo para que não haja a entrada de contratos com custo mais elevados dos mercados internacionais, priorizando o gás natural com custos mais baixos. Assim, a existência destes tipos de contratos de longo prazo induzem à valorização, levanta questões de competitividade e consequentemente impacta na formação do preço do gás natural, e esta diretriz é adotada ao gás entregue em Portugal(ERSE, 2021).
Ainda no cenário de funcionamento deste mercado, além da regulação por entidades específicas e do tipo de negociação para aquisição é importante apontar a banda de consumo, que é uma métrica que determina um limite de utilização e padroniza um tarifário para o consumidor final.

1.3 Cenário Atual Portugal e UE

Portugal, no ano de 2021, foi um dos países com mais dependência para importação de gás natural da União Europeia (UE). Segundo dados estatísticos da Direção Geral de Energia e Geologia , atualizados em 31/05/2022, o maior volume de importação ainda é assegurado pela Nigéria, que é um dos principais fornecedores de hidrocarbonetos a Portugal e mesmo com a atual guerra na Ucrânia, a exposição ao gás russo é relativamente reduzida, embora tenha crescido nos últimos anos(DGEG,2022).

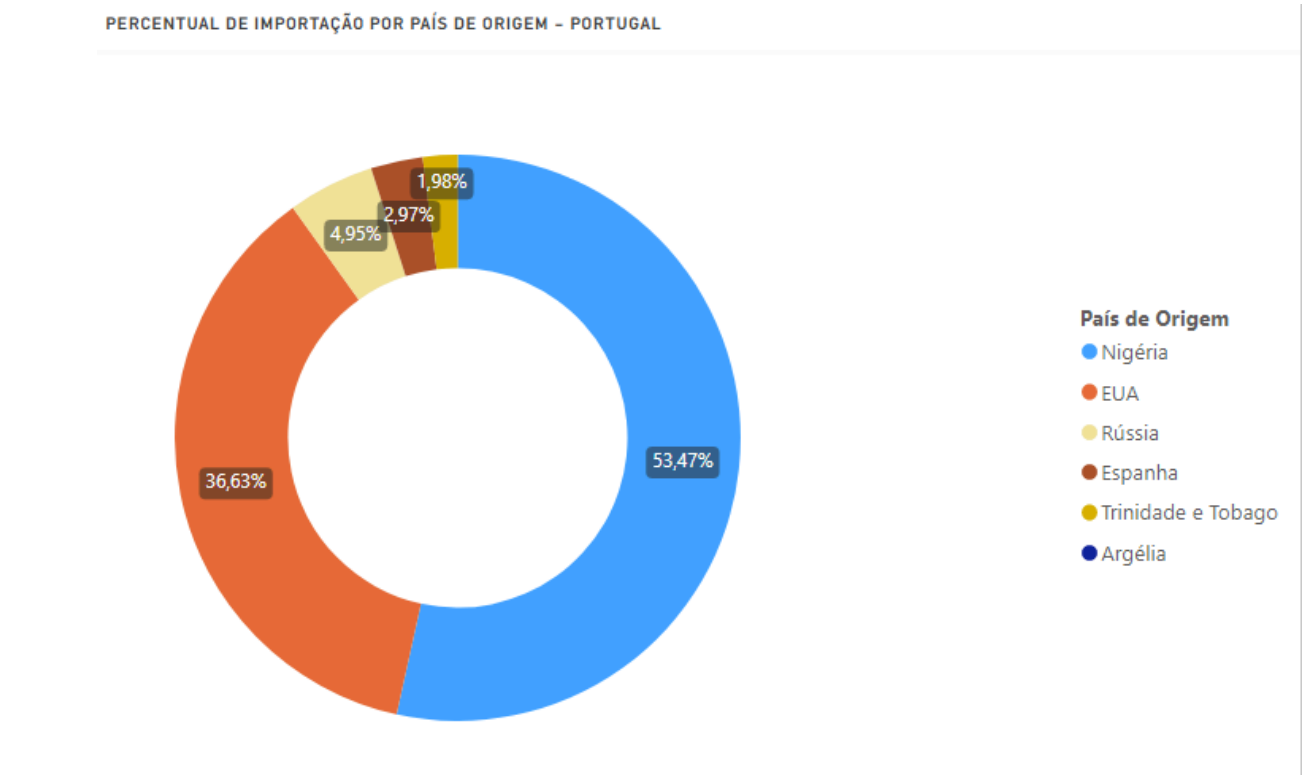

Figura 3: Percentagem de importação de gás natural, por país de origem, em Portugal
Fonte: Elaborado pelos autores - Direção Geral de Energia e Geologia (DGEG)

Em relação ao cabaz energético que corresponde à produção de eletricidade de forma sustentável, ao olhar para o mercado de calor, verificamos que a produção de aquecimento tem aumentado regularmente e quase 100% da energia primária utilizada provém do gás natural, que não é renovável. Fazendo com que a dependência de importações representem um quarto do cabaz energético do país no ano, segundo dados da Eurostat. Apesar disso, Portugal pode ser referência em fontes de energia mais limpas e rentáveis e exercer papel importante se focar no dilema de interdependência económica da União Europeia no que diz respeito às suas necessidades energéticas (PASA & BRAGHINI ,2010).
Relativo à distribuição, são utilizadas bandas de consumo como metodologias para os mercados de gás de energia, são métricas que classificam comportamentos normais típicos numa média superior a 6 meses em vez de preços fixos e caracterizadas por faixas de consumo anual:

Para residências de gás natural*:*

- Banda-D1 (Pequena): consumo anual abaixo de 20 GJ
- Banda-D2 (Médio): consumo anual entre 20 e 200 GJ
- Banda-D3 (Grande): consumo anual acima de 200 GJ

Para indústria de gás natural:

- Banda-I1: consumo anual inferior a 1.000 GJ
- Banda-I2: consumo anual entre 1.000 e 10.000 GJ
- Banda-I3: consumo anual entre 10.000 e 100.000 GJ
- Banda-I4: consumo anual entre 100.000 e 1.000.000 GJ
- Banda-I5: consumo anual entre 1.000.000 e 4.000.000 GJ
- Banda-I6: consumo anual acima de 4.000.000 GJ (voluntário)

A mais representativa segundo estudos estatísticos, é a banda D1, a referida tarifa é reguladora de venda a clientes finais.

1.4 Definir objetivos de Data Mining

Os objetivos de Data Mining deve ser visto como um processo analítico que comporte uma grande quantidade de dados numa interação sistemática através da verificação da consistência e padrões entre as variáveis escolhidas de uma determinada base de dados. Neste projecto sob um processo de mineração de dados e embasamento teórico, utiliza-se para manipulação dos dados um software que permite a partir de um conjunto de algoritmos e ferramentas a visualização e análise de dados com modelos de previsão, assim como uma linguagem de programação que proporciona interfaces gráficos e estatísticos.

Dados: Datasets importados com extensão xlsx e csv em https://ec.europa.eu/eurostat.
Processo: DCDB
Gestão de referências: Mendeley
Software : Weka
Linguagem de programação: Python

1.5 Produzir plano do projeto

Através dos conceitos de KDD (Knowledge Discovery in Databases) ou Descoberta de Conhecimento em Bases de Dados pôde-se tornar a análise dos dados compreensível e potencialmente útil, com suas etapas (Azevedo, A., & Santos, MF

,2008).

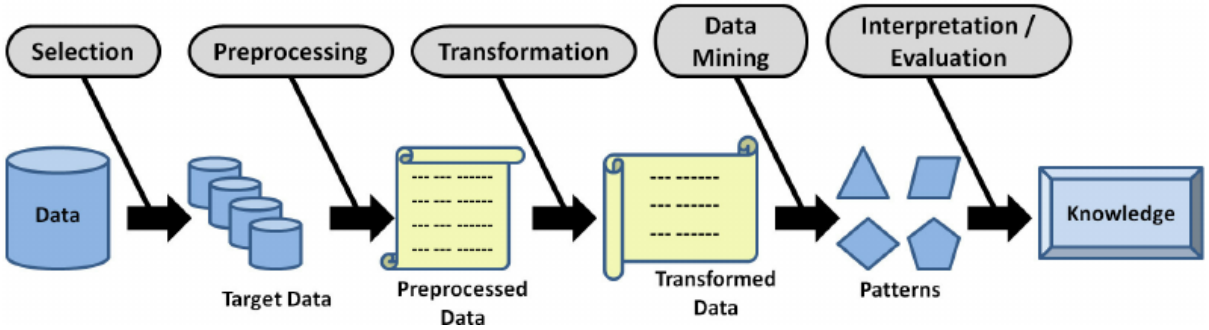

Figura 4: O processo de descoberta de conhecimento em bancos de dados (KDD)

Fonte: Elaborado por Francesco Gullo / Physics Procedia 62 (2015)

1. ***Seleção dos Dados :*** Com foco no tarifário do gás natural como objetivo tarefa, selecionou-se os conjuntos de dados com as variáveis necessárias e que impactam significativamente no critério definido.
2. ***Pré-processamento:*** Análise dos dados faltantes, limpeza e reconfiguração de horizonte temporal e formatação para consistência dos dados.
3. ***Transformação :*** Combinação dos atributos dos dados para torná-los usáveis.
4. ***Mineração de Dados :***Escolha dos algoritmos para cumprimento do objetivo tarefa e análise do comportamento dos dados.
5. ***Interpretação dos Resultados:*** Consolidação das informações e da documentação e avaliação dos padrões para resultados esperados.

2 - Estudo dos dados

2.1 Recolha dos dados iniciais

A recolha dos dados iniciais foi feita através do "Eurostat", o Serviço de Estatística da União Europeia responsável pela publicação de estatísticas e indicadores de elevada qualidade a nível europeu. Com a opção "Database" é apresentado um conjunto de tabelas disponíveis relativas ao tema "Energy".
Como o interesse inicial se prende com os preços do gás natural no consumo doméstico da banda D1, com todas as taxas e impostos associados, é necessário selecionar todas as bases de dados relativas ao tema.

Neste projeto optámos por selecionar, um total de seis bases de dados, nomeadamente:

- **"Gas prices for household consumers - bi-annual data (from 2007 onwards)"**: os preços do gás natural de consumo doméstico e não doméstico;
- **"Supply, transformation and consumption of gas - monthly data"**, mais concretamente a **"Energy balance: inlance consumption"**: o consumo interno de gás natural por país;
- **"Imports of natural gas by partner country - monthly data"**: as importações de gás natural por país;
- **"Exports of natural gas by partner country - monthly data"**: as exportações de gás natural por país**;**
- **"Cooling degree days by country - monthly data"**: o índice de grau de arrefecimento diário por país;
- **"Heating degree days by country - monthly data"**: o índice de grau de aquecimento diário por país.

2.2 Descrição dos dados

2.2.1 Preços do gás natural para consumo doméstico - dados bi-anuais (a partir de 2007)

Este *dataset* contém dados semestrais relativos aos preços do gás natural no consumo doméstico em Euro/KWH, desde o segundo semestre de 2007 ao segundo semestre de 2021, com todas as taxas incluídas, na banda de consumo D1 com o consumo anual menor que 20GJ, a mais representativa na União Europeia, inclusive Portugal, segundo o entendimento do negócio.

2.2.2 Oferta, transformação e consumo de gás natural - dados mensais

2.2.2.1 Consumo interno de gás natural

O *dataset* contém dados mensais do consumo interno de gás natural na União Europeia e em Portugal, desde janeiro de 2014 a março de 2022, em milhões de metros cúbicos.

2.2.2.2 Importações de gás natural

São disponibilizados dados mensais das importações de gás natural na União Europeia e em Portugal, desde janeiro de 2014 a março de 2022, em milhões de metros cúbicos.

2.2.2.3 Exportações de gás natural

Tal como o consumo interno e as importações, neste dataset são disponibilizados dados mensais das exportações de gás natural na União Europeia e em Portugal, desde janeiro de 2014 a março de 2022, em milhões de metros cúbicos.

2.2.3 Índices de graus de arrefecimento e aquecimento diários por país - dados mensais

2.2.3.1 Índice de grau de arrefecimento diário

Este dataset é composto por um índice do grau de arrefecimento diário (CDD), que é um índice técnico baseado no tempo concebido para descrever a necessidade dos requisitos de arrefecimento dos edifícios, que quantifica a procura por ar condicionado.

Os dias com graus de aquecimento são calculados durante um período de tempo (normalmente um ano), através da soma das diferenças entre a temperatura média diária de cada dia e a temperatura de 18°C. Para qualquer temperatura acima de 18°C, assume-se que o edifício não necessita de qualquer aquecimento. Por exemplo, três dias de Inverno consecutivos com temperaturas médias de 4°C, -2°C e -4°C totalizam 56 HDD.

É composto por dados mensais de janeiro de 1979 a dezembro de 2021 e os dados CDD são apresentados como somas de temperaturas em ˚C.

2.2.3.2 Índice de grau de aquecimento diário

O índice de graus-dia de aquecimento (HDD) é um índice técnico baseado no tempo, concebido para descrever a necessidade das condições de energia de

aquecimento dos edifícios, derivado de medições de temperatura de ar externo. Desta forma, um exemplo deste índice é o facto de quando obtemos três dias de Verão de 26°C, 28°C, e 30°C, este totalizam 30 CDD. Tal como o CDD, é composto por dados mensais de janeiro de 1979 a dezembro de 2021 e também são apresentados como somas de temperaturas em ºC (Weather Online, 2022).

2.3 Exploração dos dados

De seguida é apresentado um conjunto de estatísticas sumárias, através da linguagem de programação em Python, que permitem aprofundar as informações das variáveis disponibilizadas nos *datasets*.

2.3.1 Análise Univariada

2.3.1.1 Preços do gás natural para consumo doméstico da banda de consumo D1

De acordo com as estatísticas principais, existem 29 instâncias para Portugal e 30 instâncias para a União Europeia em termos de preços do gás natural para consumo doméstico da banda de consumo D1. A média em ambos está por volta dos 0.097 Euro/KWH, com um desvio padrão de 0.014 e 0.011 para Portugal e União Europeia, respetivamente.

Portugal

```
count    29.000000
mean      0.096279
std       0.013790
min       0.073800
25%       0.083300
50%       0.095400
75%       0.105600
max       0.125600
Name: OBS_VALUE, dtype: float64
```

União Europeia

```
count    30.000000
mean      0.097830
std       0.011063
min       0.079600
25%       0.089400
50%       0.097050
75%       0.107425
max       0.114500
Name: OBS_VALUE, dtype: float64
```

Figura 5 - Principais estatísticas sumárias dos preços do gás natural para consumo doméstico da banda de consumo D1, para Portugal e União Europeia

2.3.1.2 Consumo interno de gás natural

O consumo interno de gás natural contém 99 instâncias para Portugal e 98 para a União Europeia, com uma média mais acentuada na União Europeia devido à concentração de consumo nos 27 países, com desvio padrão maior, o que significa haver uma maior dispersão em relação aos diferentes países.

Portugal

```
count     99.000000
mean     454.495616
std       81.417860
min      281.000000
25%      386.046000
50%      462.000000
75%      519.395500
max      600.000000
Name: OBS_VALUE, dtype: float64
```

União Europeia

```
count       98.000000
mean     32543.225745
std      10727.598285
min      18431.000000
25%      22413.500000
50%      30280.293000
75%      42053.250000
max      57611.000000
Name: OBS_VALUE, dtype: float64
```

Figura 6 - Consumo interno de gás natural, para Portugal e União Europeia

2.3.1.3 Importações de gás natural

As importações de gás natural contém 99 instâncias para Portugal e 98 para a União Europeia. Portugal é um país que, atualmente, importa maioritariamente o seu consumo de gás natural, portanto apresenta-se com uma média de 474 milhões de metros cúbicos de importações de gás natural. O desvio padrão é maior na União Europeia devido às diferenças de importações entre países. Em Portugal, é também notável essa diferença, sendo relativamente menor visto se tratar apenas de um país.

Portugal

```
count     99.000000
mean     474.794515
std       99.193817
min      226.000000
25%      401.000000
50%      471.000000
75%      547.000000
max      696.000000
Name: OBS_VALUE, dtype: float64
```

União Europeia

```
count       98.000000
mean     52584.100327
std       4529.644616
min      41279.000000
25%      49143.145000
50%      52249.837500
75%      55467.667500
max      63125.410000
Name: OBS_VALUE, dtype: float64
```

Figura 7 - Importações de gás natural, para Portugal e para a União Europeia

2.3.1.4 Exportações de gás natural

O valor das exportações de gás natural é escasso para Portugal, com uma média de apenas 17 milhões de metros cúbicos, visto este ser um país que importa maioritariamente as quantidades de gás natural. Comparativamente, a União Europeia apresenta uma média de 27044 milhões de metros cúbicos em exportações de gás natural, com um desvio padrão bastante acentuado.

Portugal

```
count     96.000000
mean      17.295844
std       27.376359
min        0.000000
25%        0.000000
50%        3.000000
75%       32.250000
max      104.000000
Name: OBS_VALUE, dtype: float64
```

União Europeia

```
count       98.000000
mean     27044.010235
std       2998.717890
min      20778.655000
25%      25296.472250
50%      26636.000000
75%      28412.750000
max      37434.000000
Name: OBS_VALUE, dtype: float64
```

Figura 8 - Exportações de gás natural, para Portugal e União Europeia

2.3.1.5 Índice de grau de arrefecimento diário

Através das médias, podemos observar que, para um determinado edifício, houve uma maior necessidade de utilização de ar condicionado em Portugal comparativamente à União Europeia, devido às altas temperaturas ao longo dos anos.

Portugal

```
count    516.000000
mean      15.131822
std       26.395734
min        0.000000
25%        0.000000
50%        0.000000
75%       20.352500
max      128.980000
Name: OBS_VALUE, dtype: float64
```

União Europeia

```
count    516.000000
mean       6.089205
std       11.291459
min        0.000000
25%        0.000000
50%        0.045000
75%        6.607500
max       60.270000
Name: OBS_VALUE, dtype: float64
```

Figura 9 - CDD para Portugal e União Europeia

2.3.1.6 Índice de grau de aquecimento diário

Em Portugal, a necessidade de aquecimento em edifícios foi menor comparativamente aos resultados da União Europeia.

Portugal

```
count    516.000000
mean     103.250756
std       99.647194
min        0.000000
25%        4.710000
50%       74.170000
75%      185.912500
max      330.780000
Name: OBS_VALUE, dtype: float64
```

União Europeia

```
count    516.000000
mean     267.230233
std      187.948832
min        8.180000
25%       79.900000
50%      257.310000
75%      433.940000
max      720.410000
Name: OBS_VALUE, dtype: float64
```

Figura 10 - HDD para Portugal e União Europeia

2.4 Verificação da qualidade dos dados

Em relação ao *dataset* “Gas prices for household consumers”, em 2008, foi elaborado um relatório de auto-avaliação sobre o processo de recolha de preços de gás e electricidade. Como a metodologia para a recolha de dados sobre os preços do gás está bem definida, o conjunto de dados fornece informações sobre preços comparáveis entre países. Contudo, os utilizadores devem estar cientes de que sempre que os preços do gás em euros para países não europeus são comparados no tempo, os efeitos das diferenças nas taxas de câmbio também são tidos em conta. Sempre que os preços são calculados e/ou apresentados em euros, às taxas de câmbio médias dos dois trimestres do semestre apropriado são tomadas como referência.

De acordo com os *datasets* “Inlance consumption”, “Imports” e “Exports” o Eurostat realiza testes de qualidade, principalmente sobre a coerência da informação fornecida. Além disso, os questionários utilizados para a transmissão de dados também são incorporados em testes de coerência. Este conjunto de dados faz parte das estatísticas energéticas definidas no Regulamento (CE) n.º 1099/2008 sobre estatísticas energéticas. É considerado como estatísticas europeias e, consequentemente, aplica-se o quadro do SEE para a qualidade. Além disso, estão integrados no ciclo de Relatórios de Qualidade que tem lugar de cinco em cinco anos. Os relatórios de qualidade para as estatísticas da energia baseiam-se no artigo 6º do Regulamento (CE) nº 1099/2008 relativo às estatísticas da energia.

Finalmente, para os *datasets* “Cooling and heating degree days by country”, este conjunto de dados inclui os dados mensais publicados pelo Portal de Recursos AGRI4CAST do Centro Comum de Investigação. O Eurostat não é o produtor dos dados mensais, está apenas a re-publicá-los.

Aquando da exploração dos dados, estes estão guardados em excel e csv, de forma a proceder ao tratamento destes com a programação em Python. Foram detectados dados em falta em todos os *dataset*, pelo que teremos de proceder à preparação dos dados, nomeadamente à limpeza, construção e integração.

3 - Preparação dos Dados

Nesta etapa, os dados relativos às seis variáveis são preparados para os algoritmos de modelagem. Apesar dos dados já estarem no formato tabular, onde cada instância representa o período temporal e o país/organização, foi preciso selecionar um período de tempo e unidade de análise, tratar dos dados em falta, verificar se existem ou não *outliers* que eventualmente podem prejudicar os resultados de análise. E também fazer alterações quanto à estruturação, como por exemplo a fusão dos dados.

3.1 Seleção dos Dados

A seleção dos dados refere-se à seleção de colunas e linhas numa tabela. Não houve necessariamente necessidade de excluir colunas, no entanto, os dados relativos à União Europeia e a Portugal irão ser os mais utilizados para a análise.

Na análise de dados serão utilizadas as seguintes variáveis:

- **price** : “Gas prices for household consumers - bi-annual data (from 2007 onwards)”;
- **inlance** : “Energy balance: inlance consumption”;
- **imports** : “Imports of natural gas by partner country - monthly data”;
- **exports** : “Exports of natural gas by partner country - monthly data”;
- **cooling** : “Cooling degree days by country - monthly data”;
- **heating** : “Heating degree days by country - monthly data”.

Das seis variáveis selecionadas para análise, apenas a variável **price** tem como unidade o semestre, ao contrário das restantes cinco variáveis, que têm unidade mensal, tendo-se neste caso, optado pela realização da conversão dos dados mensais para semestrais.

Não foi necessário realizar testes de significância para decidir que campos incluir, nem recorrer a dados adicionais, uma vez que, todos os *dataset* contém a informação necessária à elaboração da análise.

3.2 Limpeza dos Dados

A limpeza dos dados é um passo importante na preparação dos dados visto que nos ajuda a garantir a qualidade dos dados.

3.2.1. Período Temporal

Os *dataset* selecionados para além de não se encontravam na mesma unidade, também não se encontravam no mesmo período de tempo, como se observa na

tabela seguinte. Só foram selecionados os dados a partir de 2014-01 até 2021-12, tendo os dados sido posteriormente convertidos para a unidade semestral.

Variáveis	Período Temporal	Unidade
price	2007 S1 a 2021 S2	Semestral
inlance	2014-01 a 2022-04	Mensal
imports	2014-01 a 2022-04	Mensal
exports	2014-01 a 2022-04	Mensal
cooling	1979-01 a 2021-12	Mensal
heating	1979-01 a 2021-12	Mensal

Tabela 1 - Período Temporal e Unidade das Variáveis

3.2.2. Dados em falta

É necessário representar os dados em falta de alguma maneira para posteriormente conseguir executar algum modelo. Uma das alternativas é a exclusão da observação/coluna, porém, este processo irá reduzir significativamente o tamanho da base de dados, comprometendo a sua viabilidade, deste modo, optou-se por substituir estes valores por zero (0) em quatro dos seis *dataset*, pois, os dados dizem respeito ao valor monetário de cada variável.

A tabela seguinte, mostra a quantidade de valores NaN existentes nos *dataset*:

Variáveis	Nº de dados NaN
price	58
inland	542
imports	565
exports	696
cooling	0
heating	0

Tabela 2 - Dados em Falta nas Variáveis

3.2.3 Detecção de Outliers

Nos dados relativos a Portugal, as variáveis **cooling** e **price** apresentam ambas outliers, em que os quartis apresentam uma diferença entre o terceiro quartil para o valor máximo.

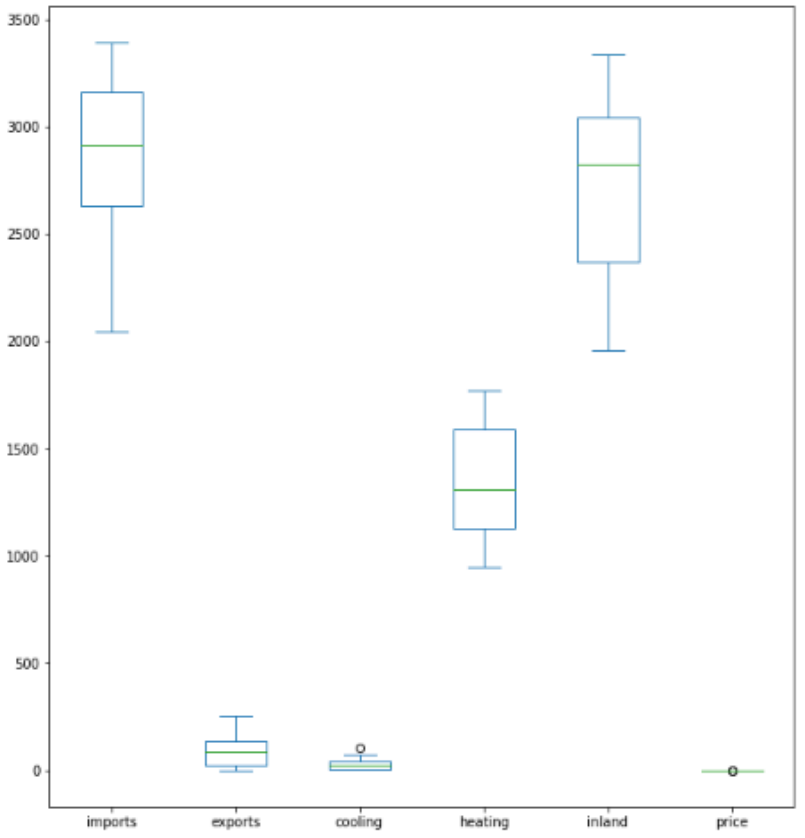

Figura 11: Box-Plot de identificação de outliers, das 6 variáveis em análise, de Portugal

cooling apresenta um outliers

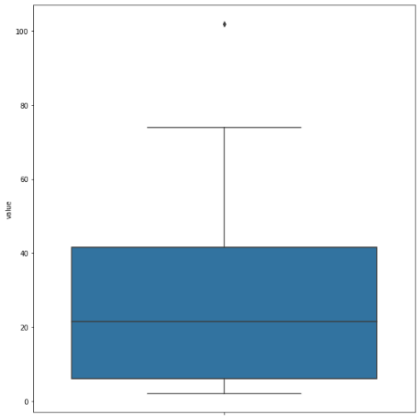

Figura 12 : Box-Plot referente à variável cooling, em Portugal

price apresenta dois outliers

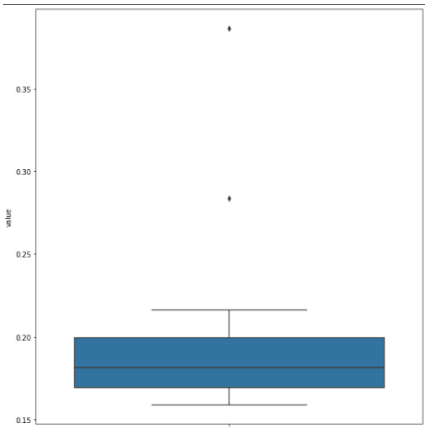

Figura 13: Box-Plot referente à variável price, em Portugal

Consoante o objetivo da análise a ser feita, a existência de outliers pode provocar o enviesamento do resultado. Neste caso, optou-se por ignorá-los, não os excluir da amostra de dados, uma vez que os dados são semestrais, o que iria afetar significativamente o tamanho da base de dados. No entanto, foi preferível tê-los em consideração, caso seja necessário realizar uma análise separada ou utilizar métodos de clusterização.

3.3 Construção e arquitetura dos Dados

Depois de harmonizar a informação constante nos diferentes *datasets* tendo em vista a utilização de uma única unidade de análise (semestre) e após o tratamento dos dados em falta, estamos em condições de proceder agrupamento das variáveis.

Uma vez que, o propósito é fazer a comparação do preço do gás entre Portugal e os países da UE, de forma a facilitar o acesso à informação, os valores referentes às seis variáveis serão agrupados numa única tabela.

O *output* a seguir apresentado, mostra a *head()* da tabela relativa aos campos e valores de Portugal que serão utilizados no modelo:

Portugal

	imports	exports	cooling	heating	inland	price
2014-S1	2186.0	253.0	5.73	1402.15	1955.0	0.1671
2014-S2	2267.0	90.0	6.05	943.57	2155.0	0.1739
2015-S1	2433.0	248.0	4.14	1600.07	2324.0	0.1703
2015-S2	2694.0	159.0	101.91	1102.99	2387.0	0.1874
2016-S1	2045.0	0.0	8.48	1581.23	2149.0	0.1651

Figura 14 - Agrupamento de variáveis

3.4 Integração e relacionamento dos Dados

Visto que cada *dataset* refere-se a uma variável, que contém os valores para a União Europeia e os países que a integram, será feita uma agregação dos dados, que consiste em criar um novo registo que resulta da informação sumária de um conjunto de registos anterior de uma tabela, de forma a simplificar os dados, serão agregados os dados da União Europeia e de Portugal numa única tabela, como se apresenta na tabela seguinte.

	União Europeia	Portugal
2014 S1	291481.0	2186.0
2014 S2	277064.0	2267.0
2015 S1	279607.0	2433.0
2015 S2	300655.0	2694.0
2016 S1	297527.0	2045.0

Figura 15 - Agregação dos dados

3.5 Formatação dos Dados

Nesta fase pretende-se explicar todas as alterações por nós efetuadas, aos dados, de modo a clarificar e torná-las mais entendíveis. No entanto, não foi necessário fazer nenhuma alteração sintática dos dados.

4 - Modelação

O objetivo principal deste projeto é a construção do modelo capaz de prever os preços do gás natural semestralmente, com base nas variáveis descritas acima.

4.1 Seleção de Técnica de Modelação

A técnica de modelação utilizada como base é a regressão linear múltipla, com base no software WEKA. O Modelo de Regressão Linear, une à equação da reta um termo aleatório para representar outros fatores, para além de variáveis independentes, que afetem a variável dependente e constitui uma das formas mais acessíveis de estabelecer a relação entre duas variáveis permitindo uma melhor análise dos dados (Curto, 2021).

4.2 Geração de desenhos e teste

Para Portugal e União Europeia, ambos os *dataset* de treino correspondem ao período temporal do primeiro semestre de 2014 até ao primeiro semestre de 2019. O *dataset* de teste é limitado ao período temporal desde o segundo semestre de 2019 até ao segundo semestre de 2021. Estes dados foram separados com base no software WEKA, na opção de *Classify*, em "*Test options*", respetivamente em "*Percentage split 70%*".

4.3 Construção do Modelo

Numa primeira fase, foi construído um modelo de regressão linear múltipla para Portugal e União Europeia, sem variáveis logaritmizadas. No software WEKA, é utilizada a regressão linear como opção de classificação que utiliza o critério AIC (*Akaike Information Criterion*) de forma a selecionar o melhor modelo. Para estes modelos, foi possível eliminar as variáveis independentes onde se observava colinearidade.

4.3.1 Regressão Linear

4.3.1.1 Portugal

Tabela X Separação dos dados teste e treino no WEKA

```
=== Run information ===

Scheme:       weka.classifiers.functions.LinearRegression -S 0 -R 1.0E-8 -additional-stats -num-decimal-places 4
Relation:     Portxlsx (1)
Instances:    16
Attributes:   6
              price
              heating
              cooling
              imports
              exports
              inland
Test mode:    split 70.0% train, remainder test
```

Tabela X Aplicação da Regressão Linear Múltipla no WEKA

```
=== Classifier model (full training set) ===


Linear Regression Model

price =

     -0.0001 * heating +
      0.0002 * cooling +
     -0.0002 * imports +
      0.0001 * exports +
      0.1897

Regression Analysis:

Variable     Coefficient     SE of Coef        t-Stat
heating          -0.0001         0.0001       -2.1741
cooling           0.0002         0.0002        1.3314
imports          -0.0002         0             -6.0904
exports           0.0001         0.0001        1.4641
const             0.1897         0.0148       12.7918

Degrees of freedom = 11
R^2 value = 0.8328
Adjusted R^2 = 0.77201
F-statistic = 13.6978

Time taken to build model: 0 seconds
```

```
=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correlation coefficient                  0.8406
Mean absolute error                      0.0066
Root mean squared error                  0.0077
Relative absolute error                 59.7778 %
Root relative squared error             54.3552 %
Total Number of Instances                5
```

Para Portugal, o algoritmo desenvolveu um modelo que seleciona todas as variáveis já existentes. O Coeficiente de Determinação (*R^2 value*) é uma medida estatística que avalia a proporção da variância na variável dependente que é prevista pela regressão linear e as variáveis independentes. O modelo apresenta um coeficiente de determinação de 0.8328, o que significa que 83,28% do preço do gás natural é explicado pelas variáveis acima apresentadas. O RMSE (*root mean squared error*): é a medida que calcula "a raiz quadrática média" dos erros entre valores

observados (reais) e predições (hipóteses). Portanto, este modelo tem uma taxa de resíduos por volta dos 60%.

Tabela X Previsão dos dados de teste, para Portugal, no WEKA

```
=== Predictions on test split ===

    inst#     actual  predicted      error
        1      0.118      0.106     -0.012
        2      0.098      0.103      0.005
        3      0.126      0.119     -0.007
        4      0.095      0.104      0.009
        5      0.096      0.096     -0
```

4.3.1.2 União Europeia

Tabela X Separação dos dados teste e treino no WEKA

```
=== Run information ===

Scheme:       weka.classifiers.functions.LinearRegression -S 0 -R 1.0E-8 -additional-stats -num-decimal-places 4
Relation:     UE (1)
Instances:    16
Attributes:   6
              price
              heating
              cooling
              imports
              exports
              inland
Test mode:    split 70.0% train, remainder test
```

Tabela X Aplicação da Regressão Linear Múltipla no WEKA

```
=== Classifier model (full training set) ===


Linear Regression Model

price =

      0.0007 * cooling +
     -0      * inland +
      0.1431

Regression Analysis:

Variable     Coefficient     SE of Coef        t-Stat
cooling           0.0007         0.0002        3.7995
inland           -0              0            -3.458
const             0.1431         0.0143        9.998

Degrees of freedom = 13
R^2 value = 0.8322
Adjusted R^2 = 0.80643
F-statistic = 32.2448

Time taken to build model: 0 seconds
```

Para a União Europeia, o algoritmo desenvolveu um modelo que seleciona apenas as variáveis independentes consumo e CDD. O Coeficiente de Determinação (*R^2 value*) é de 0.8322, o que significa que 83% do preço do gás natural é explicado pelo consumo e pelo CDD. O RMSE apresenta-se relativamente menor, comparado com Portugal.

```
Time taken to build model: 0 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correlation coefficient                  0.8175
Mean absolute error                      0.0032
Root mean squared error                  0.0037
Relative absolute error                 42.2295 %
Root relative squared error             41.8958 %
Total Number of Instances                5
```

Tabela X Previsão dos dados de teste, para a UE, no WEKA

```
=== Predictions on test split ===

    inst#     actual  predicted      error
        1      0.098      0.097     -0.001
        2      0.115      0.111     -0.004
        3      0.113      0.107     -0.006
        4      0.105      0.109      0.004
        5      0.107      0.109      0.002
```

4.4 Revisão do modelo

Em termos de qualidade do modelo de Regressão Linear, aplicado aos dados de Portugal e da União Europeia, pode-se considerar que o resultado foi previsto com precisão.

5 - Avaliação

5.1 Avaliação de resultados

Foi possível chegar a todos os objetivos previamente definidos. Considera-se que o modelo escolhido, nomeadamente Regressão Linear, foi o mais indicado, tendo apresentado uma precisão de previsão superior a 80%.

5.2 Revisão do processo

Após revisão dos processos, considerou-se que todas as tarefas relativas ao modelo foram adequadamente cumpridas.

Conclusão

As tarifas do gás natural evidentemente têm como enfoque o comercializador e consumidor final e os custos energéticos têm um peso substancial originados de diversas instâncias, principalmente pelo facto de que Portugal está inserido num mercado concorrencial. Pelo que foi constatado nos dados do Eurostat, os preços de gás natural (antes de impostos e taxas), regulados pelos escalões de consumo são diretamente proporcionais à dependência de importações.

Por processo de descoberta de Conhecimento em Bases de Dados, selecionamos as bases relativas ao consumo doméstico e não doméstico, importação , exportação e grau de aquecimento e arrefecimento, entendemos que a demanda é composta por variáveis de *gastos* onde o consumo depende da propensão ao consumo.

Relativamente às necessidades nominais de arrefecimento e aquecimento, não se aplicam às oscilações extraordinárias na comparação EU e Portugal, pois impactam em diferentes sectores de actividade e regiões, justificando o aumento ou diminuição dos preços em condições adversas. A exportação e importação, entretanto, impactam significativamente nas tarifas, pela necessidade de acomodar o aumento em volumes globais de importação (atualmente 100%) às operações e estruturas de alto consumo industrial (não doméstico) e dos consumidores finais(doméstico), elevando os custos de operação e consequentemente os preços médios gerais.

A escolha do modelo de Regressão Linear, que estuda a relação entre a variável dependente e as independentes, aparenta ser a melhor escolha para o estudo em análise, uma vez que, o coeficiente foi superior a 80%, quer para Portugal, quer para a União Europeia, tendo Portugal apresentado um valor ligeiramente mais elevado.

Referências

Azevedo, A., & Santos, MF (2008). KDD, SEMMA e CRISP-DM: uma visão paralela. IADS-DM .

DGEG. Direção Geral de Energia e Geologia. *Importações/Exportações de petróleo e derivados.* Disponível em: https://www.dgeg.gov.pt/pt/estatistica/energia/petroleo-e-derivados/importacoes-exportacoes/.

ERSE (2021). *" GÁS: COMO FUNCIONA?"*. Acedido em 29 de Junho de 2022. Disponível em : https://digital-strategy.ec.europa.eu/pt/node/11110.

EUROSTAT (2022). *Shaping Europe's digital future.* Publications Office of the European Union. Acedido em 29 de Junho de 2022. Disponível em : https://digital-strategy.ec.europa.eu/pt/node/11110.

Lopes, A. C. B. da S. (2009). *Macroeconometria I - Estimação OLS do Modelo de Regressão Linear com Séries Temporais*.

Miguel, Fábio Rafael Parreira (2010). *Metodologia de Correlação entre Consumos Energéticos e as Variáveis Influentes para Aplicação a Contratos de "Energy Services" Edifícios Saudáveis* .Faculdade de Engenharia da Universidade do Porto Mestrado Integrado em Engenharia Mecânica.

Pasa, Carine & Pasa, Leandro & Junior, Aldo & Souza, Samuel. (2012). *Evaluation of energy efficiency in buildings and their relationship with their building materials*. Revista Produção Online. 12. 10.14488/1676-1901.v12i1.873.

PORTAL DE PROCESSOS DE NEGÓCIOS DATAPREV (.*Orientação para o Processo de Desenvolvimento de Ciência de Dados e Inteligência Artificial.* PD-Dataprev: Framework Ágil de Desenvolvimento. Acedido em 19 de Julho de 2022. Disponível em: https://webcache.googleusercontent.com/search?q=cache:ModoGgGAR3sJ:desenvolvimento.dataprev.gov.br/pddataprev_internet/visualizar_guia.php?idguia%3D700&hl=pt-PT&gl=pt&strip=0&vwsrc=0

Weather Online. (2022). Acedido a 19 de julho de 2022 em https://www.weatheronline.co.uk/faq/hdd_cdd.html