

Análise de Dados na Saúde Publica Portuguesa

Rafael Antunes n°55336 Bruno Vicente n°55300 Lucas Galveias n°55053

Trabalho de Elementos de Inteligência Artificial do curso **Inteligência Artificial e Ciência de Dados** (1º ciclo de estudos)

Professor: João Carlos Raposo Neves

Maio de 2025.

Prefácio

Este trabalho prático foi realizado no âmbito da Unidade Curricular (UC) de Elementos de Inteligência Artificial e Ciência de Dados, lecionada no segundo semestre do primeiro ano da Licenciatura em Inteligência Artificial e Ciência de Dados. O principal objetivo deste trabalho é a recolha e análise de diversas estatísticas sobre Portugal, com o intuito de extrair conhecimento a partir das relações existentes entre essas estatísticas.

Resumo

Neste trabalho, foram recolhidas e analisadas estatísticas de Portugal relativas à quantidade de pessoal ao serviço nos hospitais, à taxa bruta de natalidade por mil habitantes e à taxa bruta de mortalidade por mil habitantes. O objetivo consiste em relacionar os aumentos ou decréscimos das taxas brutas de mortalidade e natalidade com as variações na quantidade de pessoal ao serviço nos hospitais.

Palavras-chave

Análise de Dados, Portugal, PORDATA, Saúde Publica, Pessoal Hospitalar, Taxa de Mortalidade, Taxa de Natalidade, Aprendizagem Não Supervisionada, Estatísticas Municipais



Índice

Pı	refác	io	iii							
Resumo										
Ín	dice		vii							
1	Intı	rodução	1							
	1.1	Contexto da Saúde Pública em Portugal	1							
	1.2	Indicadores Selecionados	1							
	1.3	Objetivos	2							
2	Rec	colha de Dados	3							
	2.1	Fonte PORDATA	3							
	2.2	Metodologia de Extração	3							
3	Inte	egração de Dados	5							
	3.1	Reestruturação dos Datasets	5							
		3.1.1 Organização de Municípios	5							
		3.1.2 Eliminação de Dados sem relevância	5							
		3.1.3 Categorização dos Profissionais Hospitalares	5							
		3.1.4 Renomeação das Colunas	6							
		3.1.5 Compreensão dos Dados 2002 e 2023	6							
		3.1.6 Conjugação dos Datasets	6							
4	Aná	álise Exploratória de Dados	9							
	4.1	Estatística Descritiva Básica	9							
	4.2	Análise de Distribuições	9							
	12	Procura por Correlações	1/1							

	4.4	Procura de Anomalias	14
	4.5	Análise de Correlações Iniciais	15
5	Lim	peza e Pré-processamento	17
	5.1	Objetivos da Limpeza	17
	5.2	Metologia de Limpeza de Outliers	17
	5.3	Pré-processamento de Dados	17
6	Aná	alise Descritiva	19
	6.1	Escolha do modelo não supervisionado	19
	6.2	Escolha da quantidade de Clusters	19
	6.3	Resultado do modelo não supervisionado K-Means	20
7	Con	nclusão	21
	7.1	Evolução do Pessoal ao Serviço nos Hospitais	21
	7.2	Análise da Taxa Bruta de Mortalidade	21
	7.3	Relação entre Pessoal Hospitalar e Taxa de Mortalidade	22
	7.4	Análise da Taxa Bruta de Natalidade e sua Relação com o Pessoal Hospitalar	22

Introdução

1.1 Contexto da Saúde Pública em Portugal

Ao longo das últimas décadas, Portugal tem registado transformações significativas em diversos setores. Entre essas mudanças, a Saúde Pública destaca-se por ser diretamente influenciada por fatores económicos, demográficos e sociais, o que torna a sua análise particularmente relevante. Neste trabalho, iremos explorar essas alterações, o seu impacto na Saúde Pública Portuguesa e por que é importante compreender essa evolução.

1.2 Indicadores Selecionados

Para obter uma visão mais abrangente da Saúde Pública em Portugal, selecionámos três indicadores principais: (1) o Número de profissionais ao serviço nos hospitais, (2) a Taxa Bruta de Mortalidade e (3) a Taxa Bruta de Natalidade. Estes indicadores foram escolhidos por refletirem, de forma direta, tanto a capacidade de resposta do sistema de saúde como as dinâmicas demográficas que influenciam a saúde coletiva do país.

- Número de profissionais ao serviço nos hospitais: Este indicador inclui médicos, enfermeiros, pessoal auxiliar, técnicos de diagnóstico e outros profissionais de saúde.
- 2. **Taxa Bruta de Mortalidade:** Expressa o número total de óbitos ocorridos por cada mil habitantes num determinado ano.
- Taxa Bruta de Natalidade: A taxa de natalidade corresponde ao número total de nascimentos ocorridos a cada mil habitantes em um determinado ano.

1.3 Objetivos

Por meio destes indicadores, pretende-se analisar a possível correlação do número de profissionais ao serviço hospitalar e o número de nascimentos e óbitos, verificando assim a sua possível influência em picos e quedas nos valores dos mesmos bem como identificar padrões, tendências e eventuais exceções que possam confirmar a existência da referida relação.

Recolha de Dados

2.1 Fonte PORDATA

O Pordata é uma plataforma online que disponibiliza um vasto conjunto de dados estatísticos organizados por diversos temas. Entre esses temas encontram-se os indicadores relevantes para a realização deste trabalho, estando os dados também estruturados por níveis geográficos, o que permite a filtragem por município, conforme solicitado. A plataforma possibilita aos utilizadores explorar essa estrutura, selecionar os indicadores de interesse e visualizar a informação sob a forma de tabelas e gráficos.

2.2 Metodologia de Extração

A metodologia adotada para a extração de dados neste trabalho baseou-se essencialmente em técnicas de *web scraping*, complementadas por etapas de processamento e estruturação da informação recolhida.

Inicialmente, foi definida uma lista de URLs correspondentes às páginas dos indicadores selecionados no portal PORDATA. Utilizou-se a biblioteca *requests* para realizar requisições *HTTP* e obter o conteúdo *HTML* de cada uma dessas páginas.

Posteriormente, esse conteúdo foi processado com o auxílio da biblioteca *BeautifulSoup*, que possibilitou a análise sintática (*parsing*) do *HTML*, convertendo-o numa estrutura hierárquica de objetos que facilitou a navegação e a extração de elementos específicos.

O principal objetivo desta etapa foi identificar e extrair os links associados ao botão de download, que redirecionavam diretamente para os *CSVs*.

Por fim, recorreu-se à biblioteca pandas para carregar estes dados extraídos diretamente para estruturas do tipo *DataFrame*, através da *função read_csv(link_download)*, permitindo assim o seu posterior tratamento e análise em ambiente *Python*.

Integração de Dados

3.1 Reestruturação dos Datasets

Neste capítulo, descrevemos em detalhe como realizamos o processo de limpeza, transformação e consolidação dos diferentes conjuntos de datasets que compõem nosso estudo. O objetivo principal foi criar uma base única e padronizada, facilitando tanto a análise exploratória quanto a análise descritiva subsequentes. A seguir, apresentamos as etapas-chave desse processo.

3.1.1 Organização de Municípios

Cada dataset apresentava uma coluna denominada "ambiente geográfico", que continha o nome do país. Esta coluna servia para distinguir os dados de âmbito nacional dos de âmbito municipal. Como o objetivo do trabalho era focar apenas nos dados municipais, foram eliminadas todas as linhas que apresentavam qualquer valor nesta coluna.

Esta filtragem teve como propósito restringir o conjunto de dados, assegurando que apenas os registos correspondentes a municípios fossem considerados, conforme exigido pelo enunciado do trabalho.

3.1.2 Eliminação de Dados sem relevância

Por predefinição, os datasets da PORDATA incluem colunas adicionais, que estão disponíveis caso se pretenda selecionar mais informações. Como não optámos por essa funcionalidade, essas colunas ficaram irrelevantes para a nossa análise, pelo que procedemos à sua eliminação.

3.1.3 Categorização dos Profissionais Hospitalares

No DataFrame dos Profissionais Hospitalares, a estrutura original apresentava uma coluna específica para cada categoria de profissional, com os respetivos totais dispostos numa coluna adjacente. Contudo, esta organização dificultava a análise e o tratamento dos dados.

Para tornar a estrutura mais legível e adequada à análise exploratória e descritiva, foi aplicada uma operação de pivot. Com esta transformação, cada categoria de profissional passou a ser representada por uma coluna distinta, contendo a respetiva quantidade de profissionais. Esta reorganização facilitou a visualização, a filtragem e a agregação dos dados nas etapas seguintes.

3.1.4 Renomeação das Colunas

As colunas dos *DataFrames* resultantes foram renomeadas para designações mais claras e intuitivas, facilitando o trabalho da conjungação dos *DataFrames* deixando assim as colunas 'Anos' e 'Distrito' estabelecidas como o índice de cada linha.

3.1.5 Compreensão dos Dados 2002 e 2023

Constatada a falta de dados para os anos de 1996 a 2001 no dataset 'Pessoal ao serviço nos hospitais (Portugal)', a análise foi restringida ao período de 2002 a 2023, uma vez que este continha dados completos para todos os anos.

3.1.6 Conjugação dos Datasets

Os DataFrames foram unificados num só através de uma operação de junção (*merge*). Esta junção foi realizada com base no índice referido anteriormente, 'Anos' e 'Distrito', utilizando uma estratégia para assegurar a inclusão de todos os registos anuais de todas as fontes, sendo 'Distrito' estabelecida como a variável geográfica única, recorrendo-se assim à eliminação da coluna referente ao município.

Sendo assim a estrutura do DataSet final ficou desta forma:

	Ano	Distrito	Nome Municipio	Tipo de Hospital	Enfermeiros	Médicos	Pessoal auxiliar	Técnicos de diagnóstico e terapêutica	Outros	Total	Taxa Bruta de Mortalidade (‰)	Taxa Bruta de Natalidade (‰)
0	2002	Distrito de Santarém	Abrantes	Hospitais	207.0	73.0	159.0	34.0	132.0	605.0	14.3	8.6
1	2002	Distrito da Guarda	Aguiar da Beira	Hospitais		0.0	0.0					7.4
2	2002	Distrito de Évora	Alandroal	Hospitais	0.0	0.0	0.0	0.0	0.0	0.0		6.0
3	2002	Distrito de Aveiro	Albergaria-a- Velha	Hospitais		0.0	0.0		0.0		10.4	11.1
4	2002	Distrito de Faro	Albufeira	Hospitais	0.0	0.0	0.0	0.0	0.0	0.0	8.8	13.7
5	2002	Distrito de Santarém	Alcanena	Hospitais		0.0	0.0		0.0			8.5
6	2002	Distrito de Leiria	Alcobaça	Hospitais	66.0	13.0		10.0	46.0	172.0		11.1
7	2002	Distrito de Setúbal	Alcochete	Hospitais		0.0	0.0		0.0			13.0
8	2002	Distrito de Faro	Alcoutim	Hospitais	0.0	0.0	0.0	0.0	0.0	0.0	25.9	5.7
9	2002	Distrito de Setúbal	Alcácer do Sal	Hospitais	0.0	0.0	0.0	0.0	0.0	0.0	12.8	7.6

Figura 3.1: Estrura do DataSet final

Análise Exploratória de Dados

4.1 Estatística Descritiva Básica

Neste capítulo começámos por fazer um resumo do DataFrame, que nos indicou o nome das colunas, seguido da quantidade de linhas não nulas. Ao compararmos este valor com o total de linhas, concluímos que não existem valores em falta.

De seguida, procurámos por linhas duplicadas, que também não existiam. Por fim, identificámos as variáveis não numéricas (Distrito) e as variáveis numéricas (Ano, Enfermeiros, Médicos, Pessoal auxiliar, Técnicos de diagnóstico e terapêutica, Outros, Total, Taxa Bruta de Mortalidade (‰) e Taxa Bruta de Natalidade (‰)).

4.2 Análise de Distribuições

Para caracterizar a forma como cada variável numérica se distribui, geraram-se histogramas e gráficos de densidade sobre os dados brutos. Permitindo identificar o perfil de distribuição de cada variável.

Primeiro, analisámos as variáveis referentes aos profissionais ao serviço nos hospitais. Observa-se uma tendência global de crescimento ao longo dos anos, com distritos mais desenvolvidos apresentando, em geral, maiores valores absolutos (ver Figura 4.1). Note-se também o salto abrupto em 2013, esse pico de uma alteração metodológica na recolha e agregação dos dados em Portugal, que passou a incluir no total distrital todos os profissionais de saúde rurais, técnicos de diagnóstico e terapeutas que antes não eram consolidados no mesmo campo.

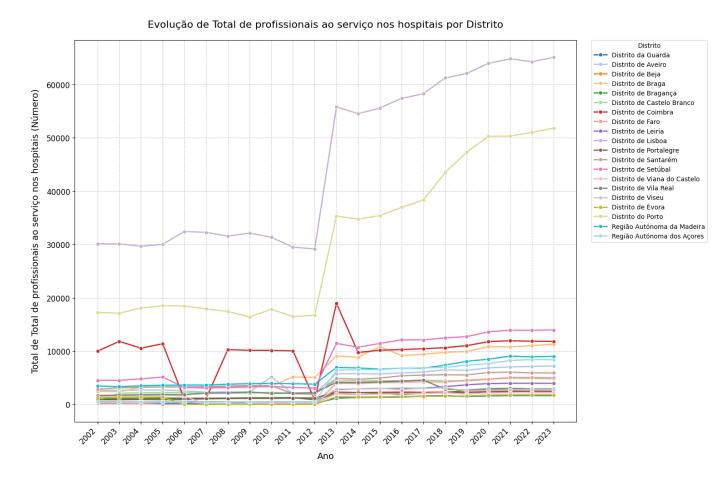


Figura 4.1: Evolução do total de profissionais ao serviço nos hospitais, por Distrito (2002–2023).

Quanto à distribuição desses valores ao longo de todos os distritos e anos, observa-se uma grande concentração em torno de dois "grupos" principais: Distritos com poucos profissionais (próximos de zero) e outros com cerca de 5 000 profissionais. Essa dispersão desigual reflete, de novo, a dispersão entre distritos menos e mais desenvolvidos (ver Figura 4.2).

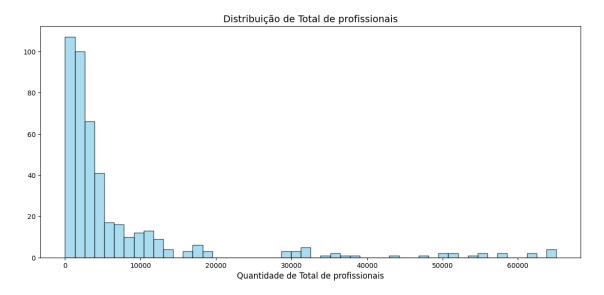


Figura 4.2: Distribuição do total de profissionais ao serviço nos hospitais, por Distrito (todos os anos).

Em seguida, examinámos a *Taxa Bruta de Mortalidade* (‰). Verifica-se uma evolução relativamente estável ao longo do período analisado, mas com distritos menos desenvolvidos a registar valores ligeiramente mais elevados em média (Figura 4.3). A distribuição dessa taxa, quando agregada por distrito e ano, aproxima-se de uma curva de distribuição normal com leve assimetria para a esquerda, conforme ilustrado na Figura 4.4.

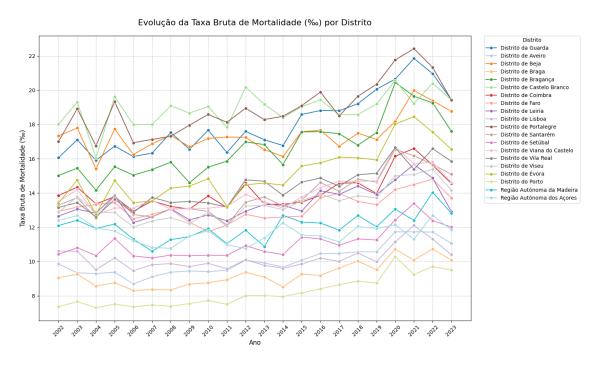


Figura 4.3: Evolução da Taxa Bruta de Mortalidade (‰), por Distrito (2002–2023).

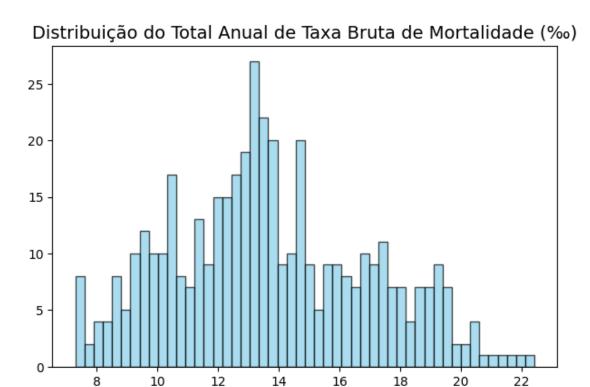


Figura 4.4: Distribuição da Taxa Bruta de Mortalidade (‰), por Distrito (todos os anos).

Quantidade de Taxa Bruta de Mortalidade (%)

Por fim, analisámos a *Taxa Bruta de Natalidade* (‰). Essa variável revela uma tendência decrescente ao longo do tempo, sem grande distinção entre distritos (Figura 4.5). A distribuição agregada das taxas de natalidade também aproxima □se de uma forma normal ligeiramente enviesada para a esquerda (Figura 4.6).

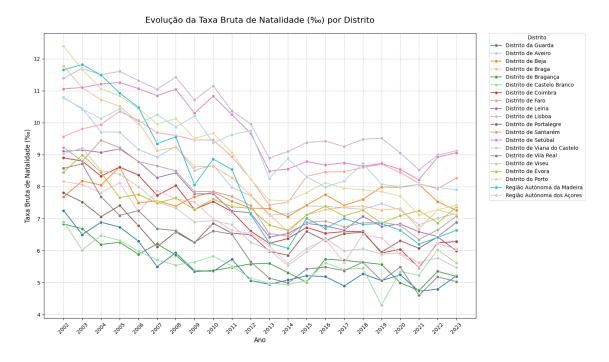


Figura 4.5: Evolução da Taxa Bruta de Natalidade (%), por Distrito (2002–2023).

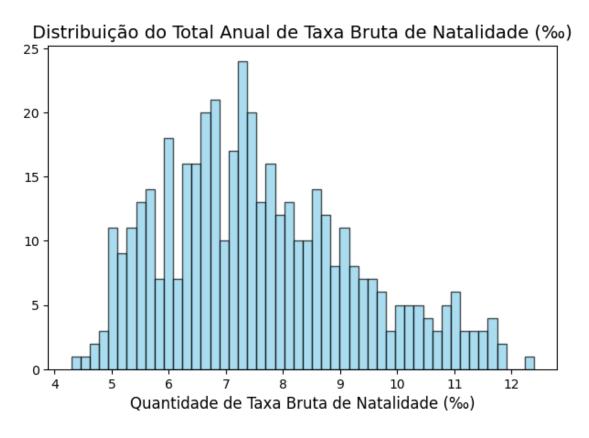


Figura 4.6: Distribuição da Taxa Bruta de Natalidade (‰), por Distrito (todos os anos).

4.3 Procura por Correlações

Adicionalmente, com o intuito de explorar possíveis relações e padrões geográficos, foram elaborados gráficos específicos para visualizar as métricas de saúde por distrito. Esta abordagem visou representar graficamente as taxas de mortalidade, natalidade e o número de profissionais em função de cada distrito. O objetivo foi facilitar a identificação de eventuais correlações ou tendências espaciais nos dados, dando o exemplo do Distrito de Castelo Branco (Figura 4.7).

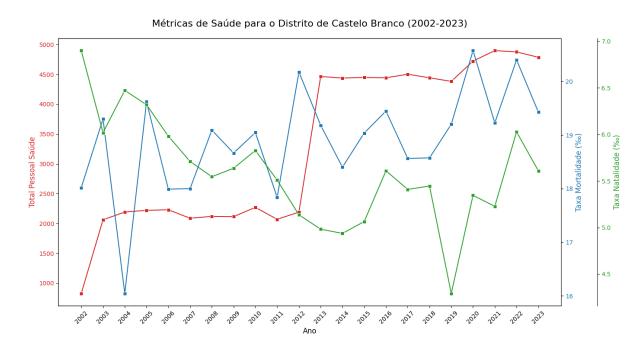


Figura 4.7: Métricas de Saude para o distrito de Castelo Branco

4.4 Procura de Anomalias

Realizando uma inspeção visual das distribuições das variáveis através dos histogramas e gráficos de densidade, forneceu informações valiosas sobre a forma e a dispersão dos dados, mas no entanto não permitiram uma identificação clara de outliers específicos.

Dada a natureza menos direta dos histogramas e gráficos de densidade para a deteção de outliers isolados, e de forma a obter uma identificação mais rigorosa, optou-se por uma abordagem estatística realizada na parte de limpeza e pré-processamento.

4.5 Análise de Correlações Iniciais

Para além dos gráficos individuais (como o da Figura 4.7) que ilustram a relação entre variáveis em cada distrito, optámos por uma abordagem global: a construção de uma matriz de correlação. Esta matriz resume estatisticamente as covariâncias normalizadas entre todos os indicadores e facilita a identificação rápida das associações mais relevantes.

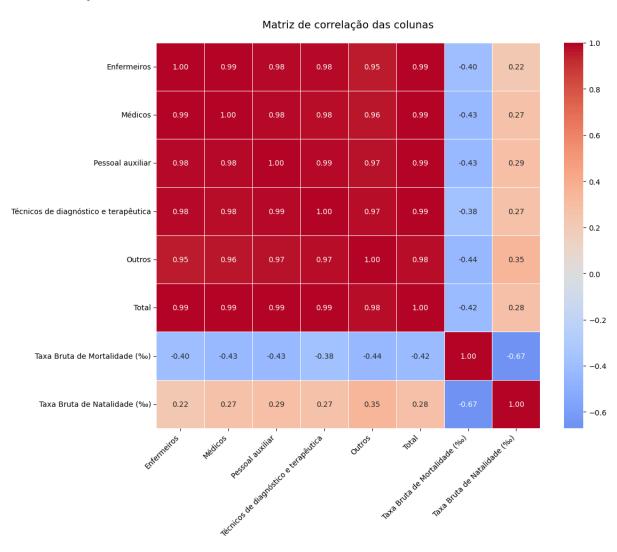


Figura 4.8: Matriz de correlação dos Indicadores

O heatmap da Figura 4.8 combina cores e anotações numéricas para mostrar a força (valor absoluto) e a direção (sinal) de cada correlação. A interpretação principal é que a quantidade de profissionais hospitalares apresenta correlação negativa moderada a forte com a Taxa Bruta de Mortalidade (‰). Ou seja, à medida que o número de profissionais hospitalares aumenta, verifica-se uma tendência de diminuição da mortalidade.

Já a correlação entre o quantidade de profissionais hospitalares e Taxa Bruta de Natalidade (‰) é muito próxima de zero, indicando praticamente nenhuma associação linear entre esses dois indicadores.

Em suma, esta análise inicial de correlações suporta a hipótese de que o reforço dos recursos humanos em saúde (profissionais hospitalares) contribui para a redução das taxas de mortalidade, mas não exerce impacto estatisticamente significativo sobre a natalidade.

Limpeza e Pré-processamento

5.1 Objetivos da Limpeza

No âmbito da Limpeza e Pré-processamento dos Dados, a análise exploratória, detalhada no capítulo anterior, revelou a ausência de valores duplicados ou em falta no conjunto de dados, não sendo necessárias intervenções a este nível. Contudo, a inspeção visual dos gráficos gerados durante a análise exploratória não permitiu uma conclusão definitiva sobre a presença de outliers.

5.2 Metologia de Limpeza de Outliers

Por isso, optou-se por recorrer a uma análise estatística com base no Z-score para identificar eventuais outliers. Com esta abordagem, detetou-se um único outlier na categoria "Outros" (pessoal ao serviço) para o distrito de Viseu, no ano de 2013.

Depois de localizar precisamente esse valor atípico, escolheu-se a técnica de Winsorização para o seu tratamento. Em vez de eliminar o ponto de dados (o que levaria à perda de informação), aplicou-se a Winsorização à coluna correspondente, ajustando o outlier ao limite do intervalo aceitável e preservando, assim, a integridade global do conjunto de dados.

5.3 Pré-processamento de Dados

Após o tratamento dos outliers, procedeu-se ao pré-processamento dos dados, etapa crucial para o bom desempenho de diversos algoritmos de Machine Learning, especialmente aqueles baseados em medidas de distância (por exemplo, o K-Means, utilizado posteriormente neste trabalho).

Para garantir que todas as variáveis contribuíssem de forma equilibrada para o treino dos modelos, estandardizou-se os dados numéricos com o StandardScaler da biblioteca scikit-learn. Este procedimento consiste na normalização de cada

variável, evitando as diferenças de escala ou de ordem de grandeza entre as variáveis.

Por fim, converteu-se a única variável categórica (distritos) numa representação numérica. Como o algoritmo K-Means não aceita valores de texto, atribuí□se a cada distrito um código inteiro de 0 a 19. Cada código corresponde a um distrito distinto, garantindo que a informação categórica pudesse ser incorporada no modelo de forma adequada.

Análise Descritiva

6.1 Escolha do modelo não supervisionado

Para a análise descritiva, foi utilizado o algoritmo K-Means, dada a sua capacidade de agrupar observações com características semelhantes de forma eficiente e interpretável. Considerando que os nossos dados contêm variáveis numéricas associadas a indicadores de saúde por distrito, o K-Means revelou-se uma ferramenta apropriada para identificar padrões e tendências gerais nos dados.

6.2 Escolha da quantidade de Clusters

Para determinar a quantidade ideal de clusters no algoritmo K-Means, utilizou-se o *método do cotovelo*, que consiste em testar vários valores de k e analisar a variação da inércia. O ponto onde ocorre a maior curvatura na curva, ou seja, onde a redução da inércia deixa de ser significativa é considerado o valor ideal. No nosso caso, a análise indicou que o valor ideal seria k=2, como mostra a Figura 6.1, valor que foi então utilizado para segmentar os dados.

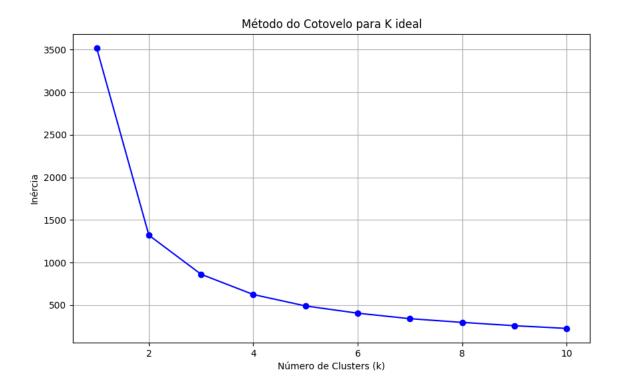


Figura 6.1: Determinação do número ideal de clusters pelo método do cotovelo

6.3 Resultado do modelo não supervisionado K-Means

Através da segmentação dos distritos com base em variáveis como o número de profissionais de saúde, a taxa de mortalidade e a taxa de natalidade, o K-Means permitiu identificar grupos com padrões semelhantes. A análise posterior destes clusters revelou uma correlação inversa entre o número de profissionais ao serviço e a taxa de mortalidade, enquanto a taxa de natalidade mostrou pouca variação entre os grupos, sugerindo uma correlação fracas. Assim, o uso do K-Means reforçou as conclusões previamente observadas na análise de correlação da Alínea 4.5.

Conclusão

As conclusões apresentadas foram fundamentadas através de uma abordagem multifacetada que incluiu a análise de séries temporais, a exploração de distribuições de dados através de gráficos de densidade, a quantificação de relações lineares por meio de uma matriz de correlação, e a identificação de agrupamentos com padrões semelhantes entre diferentes regiões ou períodos utilizando o algoritmo de clusterização K-Means.

7.1 Evolução do Pessoal ao Serviço nos Hospitais

Observa-se um crescimento no número total de profissionais de saúde ao serviço nos hospitais em Portugal ao longo do período de 2009 a 2022. No entanto, esta tendência não foi linear.

Especificamente, ao analisar a distribuição de profissionais, constata-se que, nos anos de 2011-2012, houve um decréscimo no número de profissionais. Este facto está totalmente associado ao contexto socioeconómico e às políticas de saúde da época, onde a área da saúde foi diretamente afetada por medidas que visavam a redução da despesa pública, traduzindo-se tipicamente no congelamento de admissões ou mesmo na redução de pessoal.

Após 2013, o crescimento no número de profissionais retoma e parece ser mais acentuado, especialmente a partir de 2015, continuando a tendência de aumento até 2022.

7.2 Análise da Taxa Bruta de Mortalidade

A taxa bruta de mortalidade (por 1000 habitantes) em Portugal apresenta flutuações ao longo do período analisado, mas com uma tendência geral de ligeiro aumento. Registaram-se picos nos anos de 2012 e 2015. De forma mais pronunciada, entre 2020 e 2022, observaram-se os valores mais elevados, período que coincide

claramente com o impacto da pandemia de COVID-19, com a taxa a ultrapassar valores bastante acentuados.

É importante notar que a taxa bruta de mortalidade é consideravelmente influenciada pelo envelhecimento da população. Portugal é consistentemente classificado como um dos países com uma maior proporção de idosos a nível europeu e mundial, o que contribui para uma taxa bruta de mortalidade estruturalmente mais elevada.

7.3 Relação entre Pessoal Hospitalar e Taxa de Mortalidade

Parece existir uma correlação entre a taxa bruta de mortalidade e a quantidade de profissionais ao serviço dos hospitais. Embora tanto a taxa bruta de mortalidade como o número de profissionais tenham vindo a crescer de forma geral no período mais alargado (aproximadamente de 2009 a 2022, ajustando a referência temporal para consistência com os dados detalhados), sugere-se que um maior número de profissionais poderá exercer um efeito de controlo sobre a variável da mortalidade.

A crise de 2011-2012 é um exemplo pertinente uma vez que o decréscimo no número de profissionais nesse período coincide com um crescimento da taxa de mortalidade observado na maioria dos distritos (com exceção de Beja e da Região Autónoma dos Açores). Esta observação reforça a conclusão de que a crise e os cortes na saúde afetaram a área a ponto de, possivelmente, terem contribuído para o aumento da taxa bruta de mortalidade.

Contudo, apesar do aumento geral do número de profissionais de saúde a partir de 2013 (e de forma mais expressiva depois de 2015), a taxa bruta de mortalidade nacional continuou com uma tendência de ligeiro aumento, culminando nos picos da pandemia. Isto sugere que, a nível agregado nacional, o aumento do número de profissionais, por si só, não foi suficiente para contrariar outros fatores que pressionam a taxa de mortalidade para cima, como o envelhecimento progressivo da população portuguesa e eventos de saúde pública de grande magnitude (CO-VID-19). Para isolar melhor o impacto do pessoal, seria necessário analisar taxas de mortalidade ajustadas por idade ou indicadores mais específicos de qualidade dos cuidados.

7.4 Análise da Taxa Bruta de Natalidade e sua Relação com o Pessoal Hospitalar

No que concerne à taxa bruta de natalidade, não se observa uma correlação direta e positiva entre o aumento do número de profissionais de saúde e um aumento expressivo desta taxa. A taxa de natalidade, após atingir um mínimo em anos recentes, recuperou ligeiramente, mas permaneceu em patamares historicamente baixos, mesmo com o aumento subsequente de profissionais.

Este panorama reforça a ideia de que a natalidade é primariamente influenciada por um complexo conjunto de fatores socioeconómicos, culturais e políticas familiares.