

Mestrado em Ciência de Dados
Unidade Curricular de Reconhecimento de Padrões
Docente: José Dias
Turma LOA-MCD
Ano letivo 2024/2025



Grupo composto por:
Bernardo Fleming Nº126956
Joan Concha Nº126550
Meda Racaityte Nº127575
Raquel Claudino Nº126695

Introdução	1
Caracterização dos Dados	2
PCA.....	3
Adequação PCA:.....	3
Correlações fortes e positivas:	3
Correlações fortes e negativas:	3
Relações fracas ou inexistentes	4
Número de PC.....	4
Variância Explicada	4
Critério de Kaiser	5
Gráfico do cotovelo	6
Scores PCA:.....	6
Análise dos scores:	7
Clustering	7
Hierarchical Clustering.....	8
Resultados:	8
K-Means Clustering	8
Características dos Clusters gerados:.....	8
PAM (Partitioning Around Medoids).....	9
Características dos Clusters gerados:.....	9
Gaussian Mixture Models (GMM)	9
Integração das Variáveis Profile	10
Análise dos Clusters.....	11
Cluster 1: Trabalho árduo	11
Cluster 2: Europa tradicional	12
Cluster 3: Emergência Oriental e de Portugal	13
Cluster 4: Centros Urbanos e Capitais Económicas	14
Conclusão.....	16
Bibliografia	17
Anexo A	18
Preparação dos Dados	18
Recolha e Tratamento	18
Análise Exploratória	19
Imputação de Valores Omissos e Análise de Outliers	23
Anexo B	35
PCA	35
Pair plot 1:	35
Pair plot 2:	36
Pair plot 3:	37
Pair plot 4:	38

Visualização da correlação:	39
Matriz de correlação:	39
Teste de Bartlett e KMO:	41
Número de PC:	42
Variância explicada:	45
Comunalidades:	46
Extração dos componentes principais:	47
Rotação:	48
Scores:	52
PC1 vs PC2.....	54
Anexo C	55
Clustering.....	55
Hierarchical Clustering:	55
K-Means:	57
PAM:	57
Gaussian Mixture Model:	58
Variáveis de Profile em comparação aos Clusters formados pelo K-Means:	60
O mistério da Andaluzia.....	61

Introdução

Este projeto foi desenvolvido no âmbito de um desafio apresentado pelo nosso cliente, a Comissão Europeia.

O desafio consiste em identificar *clusters* de regiões ao nível NUTS 2 com base em indicadores relacionados com a empregabilidade, como taxas de desemprego, empregabilidade em diferentes sectores e outros indicadores relacionados com este tema, organizados por níveis de educação, sexo, idade, entre outros.

Esta análise e a identificação de padrões regionais têm uma relevância prática significativa, pois os resultados podem ajudar na alocação mais eficiente de recursos financeiros, como fundos europeus de coesão, e apoiar a criação de políticas públicas específicas para cada região.

Para a realização deste projeto, utilizámos técnicas de análise multivariada, como Análise de Componentes Principais (PCA), para reduzir a dimensionalidade dos dados, e análise de *Clustering* para agrupar regiões com características socioeconómicas semelhantes.

Através desta análise foi possível adquirir conhecimentos essenciais para concretizar o desafio proposto pelo nosso cliente.

Caracterização dos Dados

A base de dados criada para a concretização desta análise é constituída por 51 (cinquenta e uma) variáveis ativas (*INPUT*) que incluem dados sobre diversos componentes associados à empregabilidade tais como: Empregabilidade por Indústria, Taxa de Desemprego, Compensação Monetária de Funcionários, Média de Horas Semanais de Trabalho, Recursos Humanos em Ciências e Tecnologias entre outros. Estes componentes também são subcategorizados consoante outras características como sexo, idade e nível de educação. Todas as variáveis ativas da base de dados são numéricas, com exceção do código de cada região.

As variáveis passivas (*PROFILE*) que integram esta base de dados são PIB ao Preço do Mercado Corrente (métrica), País (nominal), Percentagem de Grau de Urbanização (métrica) e a Capital Nacional (binária).

Nome Var. INPUT	Definição da Variável	Métrica
EmpGap_G	Diferença de Empregabilidade entre Géneros	Diferença de empregabilidade (%)
Emp_HTech	Emprego nos setores de Alta Tecnologia	Taxa de emprego (%)
Emp_Hrs	Horas de trabalho	Tempo (milhares de h)
Emp_Hrs_Agr	Horas de trabalho (setores Agricultura, Silvicultura e Pesca)	Tempo (milhares de h)
Emp_Hrs_Ind	Horas de trabalho (setor Indústria, exceto Construção)	Tempo (milhares de h)
Emp_Hrs_Man	Horas de trabalho (setor Indústria Transformadora)	Tempo (milhares de h)
Emp_Hrs_Cons	Horas de trabalho (setor Construção)	Tempo (milhares de h)
Emp_Hrs_TS	Horas de trabalho (setores Comércio, Transportes e Alojamento)	Tempo (milhares de h)
Emp_Hrs_Fin	Horas de trabalho (setores Atividades Financeiras e Imobiliárias)	Tempo (milhares de h)
Emp_Hrs_Pub	Horas de trabalho (setores Administração Pública e Serviços Sociais)	Tempo (milhares de h)
EmpRate_15_64	Taxa de emprego (faixa etária dos 15 aos 64 anos)	Taxa de emprego (%)
UnempRate	Taxa de desemprego	Taxa de desemprego (%)
CompEmp_Euro	Compensação por funcionário e horas trabalhadas por pessoa empregada	Compensação (€)
CompEmp	Compensação de funcionários	Compensação (€)
CompEmp_Agr	Compensação de funcionários (setores Agricultura, Silvicultura e Pesca)	Compensação (€)
CompEmp_Ind	Compensação de funcionários (setor Indústria, exceto Construção)	Compensação (€)
CompEmp_Man	Compensação de funcionários (setor Indústria Transformadora)	Compensação (€)
CompEmp_Cons	Compensação de funcionários (setor Construção)	Compensação (€)
CompEmp_TS	Compensação de funcionários (setores Comércio, Transportes e Alojamento)	Compensação (€)
CompEmp_Fin	Compensação de funcionários (setores Atividades Financeiras e Imobiliárias)	Compensação (€)
CompEmp_Pub	Compensação de funcionários (setor Administração Pública e Serviços Sociais)	Compensação (€)
Cult_Emp	Emprego no setor cultural	População (milhares)
HR_Tech	Recursos humanos em Ciência e Tecnologia (HRST)	Taxa de população (%)
RL_prod	Produtividade real do trabalho	Produtividade real (%)
Y_EmpRate_M	Taxa de empregabilidade jovem (Masculino)	Taxa de emprego (%)
Y_EmpRate_F	Taxa de empregabilidade jovem (Feminino)	Taxa de emprego (%)
Y_EmpRate_Age	Taxa de emprego jovem por faixa etária	Taxa de emprego (%)
Emp_15_64	Emprego na população ativa (15-64 anos)	População (milhares)
Emp_65p	Emprego na população sénior (65+ anos)	População (milhares)
Emp_M	Emprego masculino	População (milhares)
Emp_F	Emprego feminino	População (milhares)
PopEmp_15p	População empregada (15+ anos)	População (milhares)
PopUnemp_15p	População desempregada (15+ anos)	População (milhares)
Emp_15_64_LE	Emprego na população ativa (15-64 anos) com educação básica	População (milhares)
Emp_15_64_ME	Emprego na população ativa (15-64 anos) com educação secundária	População (milhares)
Emp_15_64_HE	Emprego na população ativa (15-64 anos) com educação superior	População (milhares)
EA_M_15_64	População masculina economicamente ativa (15-64 anos)	População (milhares)
EA_F_15_64	População feminina economicamente ativa (15-64 anos)	População (milhares)
EA_Total_65p	População sénior economicamente ativa (65+ anos)	População (milhares)
Emp_Tech_Total	Emprego em setores de tecnologia e conhecimento intensivo	População (milhares)
AWH_M_15_64	Média de horas semanais trabalhadas por homens (15-64 anos)	Tempo (h)
AWH_F_15_64	Média de horas semanais trabalhadas por mulheres (15-64 anos)	Tempo (h)
AWH_Total_65p	Média de horas semanais trabalhadas por seniores (65+ anos)	Tempo (h)
Emp_FT_M_15_64	Emprego masculino a tempo inteiro na população economicamente ativa (15-64 anos)	População (milhares)
Emp_FT_F_15_64	Emprego feminino a tempo inteiro na população economicamente ativa (15-64 anos)	População (milhares)
Emp_PT_M_15_64	Emprego masculino a tempo parcial na população economicamente ativa (15-64 anos)	População (milhares)
Emp_PT_F_15_64	Emprego feminino a tempo parcial na população economicamente ativa (15-64 anos)	População (milhares)
NEET_M	Taxa de jovens do sexo masculino que não estuda nem trabalha (NEET)	Taxa NEET (%)
NEET_F	Taxa de jovens do sexo feminino que não estuda nem trabalha (NEET)	Taxa NEET (%)
Unemp_CZ	Desemprego de cidadãos nacionais	População (milhares)
LT_Unemp_Age	Desemprego de longa duração (12 meses ou mais) na população ativa (15-64 anos)	População (milhares)

Figura 1 - Detalhes da Variáveis INPUT

PCA

A Análise de Componentes Principais (PCA) é um método estatístico utilizado para reduzir a dimensionalidade de conjuntos de dados, mantendo ao máximo a variabilidade original.

Adequação PCA:

Para avaliar a adequação do PCA, uma análise preliminar das correlações entre as variáveis é essencial. A matriz de correlação ajuda a identificar quais variáveis estão fortemente relacionadas e, portanto, são candidatas ideais para serem resumidas por componentes principais. A matriz de correlação contém coeficientes que medem a relação linear entre diferentes variáveis. Os coeficientes variam de -1 (correlação negativa perfeita) a 1 (correlação positiva perfeita), com 0 indicando a ausência de correlação linear.

Correlações fortes e positivas:

- “Emp_Hrs_Ind” e “Emp_Hrs_Man” apresentam uma correlação extremamente forte (0,995), refletindo a interdependência entre os setores industrial e de fabrico.
- “Emp_Hrs_Fin” e “Emp_Hrs_Pub” têm uma correlação muito alta (0,940), indicando que regiões com maior foco em serviços financeiros tendem a ter também um setor público mais ativo.
- As compensações em diferentes setores (“CompEmp_”) mostram fortes correlações, como entre “CompEmp_TS” (serviços técnicos) e “CompEmp_Fin” (serviços financeiros) (0,971), sugerindo que mudanças salariais num setor podem influenciar outros.
- “Emp_15_64_HE” (emprego com educação superior) tem uma forte correlação positiva com o emprego total na faixa etária de 15 a 64 anos (0,920), destacando o impacto positivo da educação superior no mercado de trabalho.
- As taxas de emprego masculino (“EA_M_15_64”) e feminino (“EA_F_15_64”) apresentam uma correlação muito alta (0,989), indicando que regiões com altos índices de emprego masculino tendem a ter também altos índices de emprego feminino. A mesma tendência observa-se também para as taxas NEET (0,752).

Correlações fortes e negativas:

- A variável “EmpRate_15_64” (taxa de emprego) tem uma correlação negativa forte com “UnempRate” (taxa de desemprego) (-0,840). Este é um resultado esperado, pois regiões com maior taxa de emprego apresentam menor desemprego.
- As taxas NEET masculina (“NEET_M”) e feminina (“NEET_F”) têm correlações negativas significativas com as taxas de emprego, -0,878 e -0,802 respectivamente.

Isso sugere que regiões com maior proporção de jovens fora do mercado têm piores indicadores gerais de emprego.

- A variável “HR_Tech” apresenta uma correlação negativa moderada com a lacuna geral de emprego (“EmpGap_G”) (-0,620), indicando que regiões com mais profissionais de recursos humanos no setor de tecnologia tendem a ter lacunas menores no mercado laboral.

Relações fracas ou inexistentes

- A produtividade regional (“RL_prod”) tem relações fracas ou inexistentes com a maioria das variáveis relacionadas ao emprego, como:
 - “RL_prod” vs “Emp_Hrs” com 0,101.
 - “RL_prod” vs “CompEmp” com -0,093.

Isso pode indicar que a produtividade não está diretamente associada ao volume total de horas trabalhadas ou compensações salariais.

- As horas médias trabalhadas por homens (“AWH_M_15_64”) e mulheres (“AWH_F_15_64”) têm correlações fracas com variáveis como taxa geral de emprego:
 - “AWH_M_15_64” vs “EmpRate_15_64” com -0,284.
 - “AWH_F_15_64” vs “EmpRate_15_64” com -0,324.

Isso sugere que o número médio de horas trabalhadas não é um determinante direto das taxas gerais de emprego.

- O setor agrícola apresenta relações fracas com outros setores, refletindo a especialização do setor agrícola em algumas regiões, sem grande interação com outros setores. Por exemplo: “Emp_Hrs_Agr” vs “Emp_Hrs_TS” com 0,239.

Número de PC

Variância Explicada

A variância explicada quantifica a proporção da variabilidade total num conjunto de dados que é capturada pelos componentes principais. Deste modo, a variância explicada mede o quanto da variação nos dados pode ser atribuída aos componentes considerados no modelo PCA.

- PC1 (Primeiro Componente Principal) explica 52% da variância total, indicando que consegue capturar a maior parte da informação nos dados. Este componente é fortemente associado às variáveis relacionadas às horas trabalhadas (“Emp_Hrs”, “Emp_Hrs_TS”, “Emp_Hrs_Cons”, “Emp_Hrs_Fin”), sugerindo que reflete fatores relacionados ao emprego e carga horária.

- PC2 (Segundo Componente Principal) explica 18% da variância, elevando o total acumulado de variância explicada para 70%. Este componente está negativamente associado à taxa de emprego (“EmpRate_15_64”) e positivamente relacionada à taxa de desemprego (“UnempRate”), refletindo um contraste claro entre emprego e desemprego. Além disso, apresenta correlações positivas significativas com variáveis como “NEET_M” e “NEET_F”, indicando uma ligação com jovens fora do mercado de trabalho ou educação.
- PC3 (Terceiro Componente Principal) contribui com 8% da variância, relacionado a setores industriais e de fabrico (“Emp_Hrs_Ind”, “Emp_Hrs_Man”).
- PC4 (Quarto Componente Principal) explica 5% da variância, capturando relações específicas entre setores como tecnologia (“HR_Tech” com correlação positiva) e agricultura (“CompEmp_Agr” com correlação negativa).
- PC5 (Quinto Componente Principal) explica 3% da variância, representando variações menores nos setores específicos, como construção (“CompEmp_Cons”) e serviços técnicos (“CompEmp_TS”), além de estar moderadamente associada a padrões de horas trabalhadas por género.

Assim, os cinco primeiros componentes juntos explicam 86% da variância total, o que é satisfatório para capturar a maior parte da informação contida nos dados originais. Componentes adicionais explicam proporções muito pequenas de variância ($\leq 2\%$) e são menos relevantes.

Critério de Kaiser

O critério de Kaiser é um método utilizado para determinar o número adequado de componentes a serem retidos. Este critério sugere que apenas os componentes principais com variância superior a 1 devem ser considerados significativos e, portanto, retidos na análise.

Observando a variância dos componentes principais, na figura identificou-se que apenas os primeiros cinco componentes principais apresentam uma variância superior a 1. Assim, pelo critério de Kaiser, o número ótimo é 5.

```
[1] 26.589 9.041 3.912 2.445 1.735 0.978 0.951 0.800
```

Figura 2 - Critério de Kaiser

Gráfico do cotovelo

O gráfico do cotovelo é uma ferramenta visual utilizada para ajudar a determinar o número apropriado de componentes a serem retidos. Este gráfico representa a variância dos componentes principais em ordem decrescente. O ponto onde a curva começa a estabilizar-se é chamado de "cotovelo". Este é o ponto onde a adição de mais componentes não contribui significativamente para explicar mais variância nos dados.

No gráfico do cotovelo nota-se uma linearidade a partir do sexto componente principal. Assim, depois do quinto componente, já não se adiciona variância explicada relevante para ser necessário incluir o sexto.

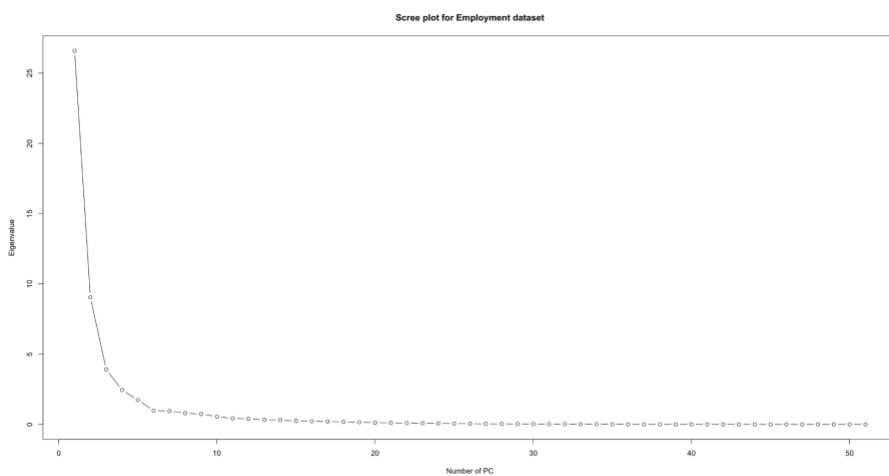


Figura 3 - Gráfico do cotovelo

Scores PCA:

Os scores dos componentes principais fornecem os *insights* sobre como diferentes observações se posicionam em relação às dimensões principais identificadas na análise.

PC1 - Dimensão Geral do Emprego:

- Scores positivos indicam regiões com altos níveis de emprego e participação no mercado de trabalho.
- Scores negativos sugerem menor participação no emprego.

PC2 - Faces do Mercado de Trabalho:

- Scores positivos indicam melhores taxas de emprego.
- Scores negativos estão associados a taxas mais altas de desemprego.

PC3 - Emprego industrial:

- Scores positivos indicam uma maior presença ou dependência de setores industriais/fábrica.
- Scores negativos sugerem menor foco nesses setores.

PC4 - Inovação vs Tradição:

- Scores positivos sugerem uma ênfase maior em tecnologia.
- Scores negativos indicam uma dependência maior da agricultura.

PC5 - Raízes do Desemprego:

- Scores positivos podem indicar regiões afetadas por desemprego estrutural ou de longo prazo.
- Scores negativos sugerem menor impacto do desemprego estrutural.

Análise dos scores:

- **Dimensão Geral do Emprego:** A maioria das observações está ligeiramente abaixo da média, sugerindo níveis de emprego moderados ou baixos.
- **Fases do Mercado de Trabalho:** As observações tendem a estar associadas a taxas de desemprego mais altas, dado o predomínio de scores negativos.
- **Emprego Industrial:** Todas as observações têm scores negativos, indicando menor presença ou dependência de setores industriais/fábrica.
- **Inovação vs Tradição:** A maioria das observações tem scores positivos, sugerindo um foco maior em tecnologia em comparação com agricultura.
- **Raízes do Desemprego:** A maioria das observações também apresenta scores positivos, indicando alguma exposição ao desemprego estrutural.

Clustering

O *clustering* é uma técnica de aprendizagem não supervisionada cujo objetivo é agrupar observações em subconjuntos, ou *clusters*, com base em semelhanças entre elas. Ao contrário de métodos supervisionados, o *clustering* não utiliza rótulos predefinidos, mas busca padrões implícitos nos dados, agrupando observações que compartilham características semelhantes, nem sempre óbvias. Essa abordagem é amplamente utilizada em análises socioeconómicas, pois permite identificar grupos de regiões ou indivíduos com desafios, ou atributos em comum.

No contexto desta análise, após a redução de dimensionalidade e seleção dos componentes principais (*Principal Components* - PCs), diferentes técnicas de *clustering* foram aplicadas visando agrupar regiões NUTS 2 com base em similaridades socioeconómicas capturadas pelas PCs. Foram avaliados quatro métodos de *clustering*: *Hierarchical*, *K-Means*, *PAM* (*Partitioning Around Medoids*) e *Gaussian Mixture*. Para avaliar a qualidade de agrupamentos (*clusters*) criados pelos métodos, foi utilizado a métrica de silhueta. Ela mede a consistência de cada observação relativamente ao *cluster* ao qual pertence, calculando-se a média das distâncias de para todas as outras observações dentro do mesmo *cluster* e a

média das distâncias de para todas as observações no agrupamento mais próximo. Os valores variam de -1 a 1 e fornecem uma métrica global da qualidade do *clustering*.

Hierarchical Clustering

- Agrupa os *Clusters* de maneira a deixar regiões semelhantes mais próximas entre si numa estrutura hierárquica.
- Método hierárquico utilizado: Ward.D2, que procura minimizar a variância das iterações dentro do *cluster*.
- Número de *Clusters*: 3, definidos visualmente no dendrograma.
- Desempenho: A avaliação de desempenho, feita através da técnica de silhueta, apresentou heterogeneidade moderada, com observações próximas às bordas dos agrupamentos.

Resultados:

Os *clusters* apresentaram os seguintes perfis médios relativamente às dimensões analisadas:

- *Cluster 1*: Baixa industrialização e inovação.
- *Cluster 2*: Níveis elevados de emprego industrial e inovação.
- *Cluster 3*: Contribuição alta do desemprego.

K-Means Clustering

- Agrupa valores em torno de um centro comum ao conjunto.
- Número de *Clusters*: 4, escolhido com base no método do cotovelo e confirmado pelo coeficiente da silhueta.
- Desempenho: Médio de 0,3, com destaque para o *Cluster 3* (0,42) indicando boa coesão interna. Ver anexo C.

Características dos Clusters gerados:

- *Cluster 1*: Grande heterogeneidade, com áreas rurais ou menos urbanizadas e regiões ricas e industrializadas.

- *Cluster 2*: Regiões urbanas intermediárias, forte presença em países ocidentais desenvolvidos.
- *Cluster 3*: Regiões menos urbanizadas e com PIB mais baixo, concentrado em países da Europa Central e do Leste.
- *Cluster 4*: Maior urbanização, inclui a maioria das capitais, representa centros econômicos desenvolvidos

PAM (Partitioning Around Medoids)

- Cria agrupamentos baseado em um elemento central comum ao conjunto.
- Número de *Clusters*: 4, alinhado ao resultado do *K-Means*.
- Desempenho: Fraco de 0,22, consideravelmente pior que o *K-Means*, com o *Cluster 2* apresentando o maior valor (0,31). Ver anexo C.

Características dos *Clusters* gerados:

- Similares aos clusters gerados no *K-Means*.
- Transferência de algumas regiões do *Cluster 2* para o *Cluster 4*.

Gaussian Mixture Models (GMM)

- Tenta estimar não só o centro dos *Clusters*, mas também a sua forma.
- Número de *Clusters*: 6, definido pelo próprio modelo.
- Desempenho: Muito baixo, o coeficiente de silhueta média foi de 0,07, com destaque negativo para os *Clusters 5* e *6* (por ficarem com valores negativos). Ver anexo C.

Conforme descrito acima, entre os métodos avaliados, o *K-Means* apresentou a melhor média da métrica da *silhueta* (0,3), o que reflete uma moderada sobreposição entre *clusters*, o que é esperado em dados socioeconómicos devido à diversidade e complexidade das regiões analisadas. O método funciona iterativamente para dividir o conjunto de dados em grupos, minimizando a variabilidade interna dos *clusters* (a soma dos quadrados das distâncias entre os pontos e o centroide do *cluster*).

Integração das Variáveis *Profile*

As variáveis passivas (PROFILE) foram utilizadas para caracterizar cada um dos *clusters* de forma mais detalhada. Estas variáveis incluem:

- PIB ao Preço do Mercado Corrente (métrica): Representa o desempenho económico das regiões em termos absolutos.
- País (nominal): Identifica a origem nacional das regiões, permitindo capturar influências culturais ou políticas.
- Percentagem de Grau de Urbanização (métrica): Reflete a proporção da população vivendo em áreas urbanas, um indicador de desenvolvimento e infraestrutura.
- Capital Nacional (binária): Indica se a região contém a capital do país, frequentemente associada a centros administrativos e económicos.

Caracterização detalhada dos clusters com as variáveis de *Profile*:

- Urbanização: *Cluster 4* apresentou maior média de urbanização, enquanto o *Cluster 3* teve os menores valores médios.
- PIB: O *Cluster 4* destacou-se com o maior PIB médio, refletindo a presença de regiões urbanizadas e capitais. O *Cluster 2*, por outro lado, apresentou menor PIB médio.
- Capitais: Pode-se observar que o *Cluster 4* concentrou a maioria das capitais, incluindo Paris (FR10), Berlim (DE30) e Lisboa (PT17), destacando a sua relevância económica e política dessas regiões. *Clusters 2* e *3* também contêm capitais, mas em menor quantidade.
- Distribuição por Países: A distribuição dos *clusters* por países é heterogénea, como mostrado no gráfico de percentagem de regiões por *cluster*. Alguns *insights* incluem:
 - Nenhum país tem regiões representadas em todos os *Clusters*, o que indica que os *clusters* desenhados conseguem captar características socioeconómicas a nível nacional de cada membro da UE.
 - Regiões do leste europeu (ex.: Romênia e Bulgária) são predominantemente classificadas no *Cluster 3*, refletindo desafios socioeconómicos maiores.
 - Regiões nórdicas (ex.: Dinamarca e Suécia) aparecem em *Clusters 2* e *4*, indicando melhor desempenho económico.

Análise dos Clusters

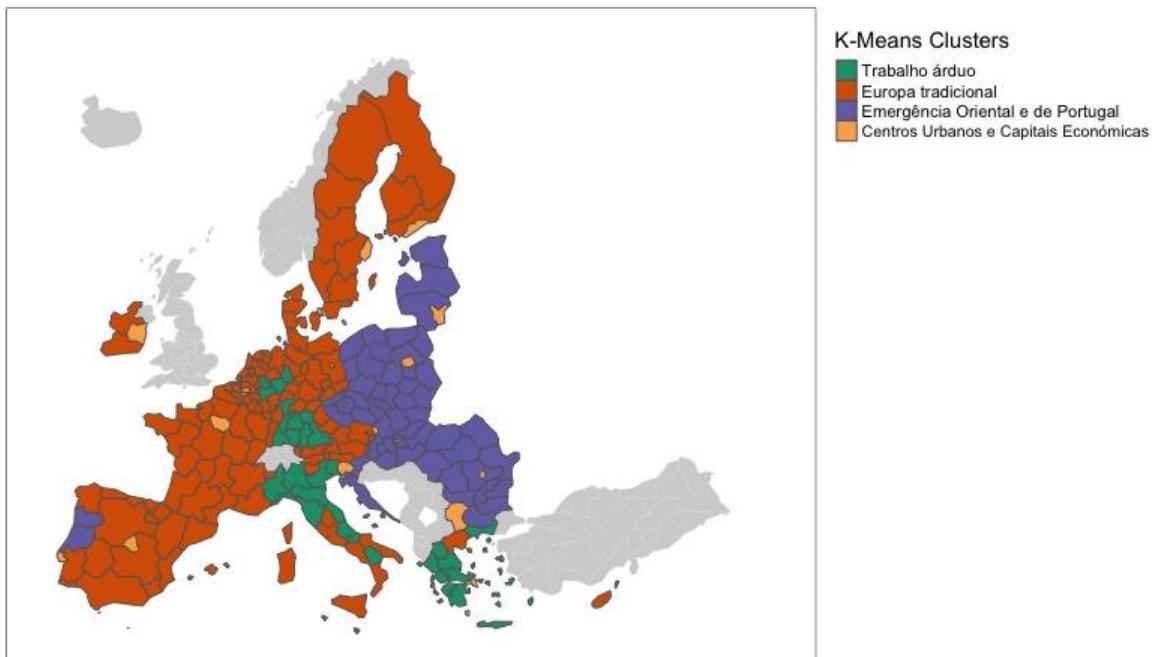


Figura 4 – K-Means Clusters

Cluster 1: Trabalho árduo

Descrição:

Esse *Cluster* possui regiões que aparentemente são distintas entre si, combinando regiões industriais ricas com regiões agrícolas de menor produtividade. A característica principal desse agrupamento é a disposição para o trabalho. São regiões que apresentam algumas das maiores quantidades de horas trabalhadas por semana, característica muito importante para o PC “Dimensão Geral do Emprego”. Ainda que a produtividade e o rendimento variem entre elas, e os setores serem muito díspares, essa importante característica criou um núcleo comum entre todas. É o único *Cluster* que não contém uma capital nacional.

Características Comuns:

- Altos valores de emprego total e número de horas trabalhadas.

Exemplos de Regiões:

- Alemanha: Regiões industriais ricas como Estugarda (DE11).
- Itália: Piemonte (ITC1), tem altos níveis de emprego industrial e tecnológico.
- Grécia: Todas as regiões menos a região de Atenas (capital). Setores tradicionais, como agricultura e serviços, com menor produtividade.

Lista de regiões abrangidas:

"DE11", "DE12", "DE13", "DE14", "DE21", "DE25", "DE27", "DE71", "DEA1", "DEA2", "DEA4", "DEA5", "EL41", "EL42", "EL43", "EL51", "EL53", "EL54", "EL61", "EL62", "EL63", "EL64", "EL65", "ES63", "ES64", "FRY3", "FRY5", "ITC1", "ITC3", "ITC4", "ITF1", "ITF2", "ITF5", "ITH3", "ITH4", "ITH5", "ITI1", "ITI3".

Cluster 2: Europa tradicional

Descrição:

Esse *cluster* agrupa o maior número de regiões. Essas possuem um PIB intermediário e urbanização moderada. Estas áreas possuem economias mais diversificadas, com uma mescla de setores industriais, agrícolas e de serviços. Inclui muitas regiões economicamente desenvolvidas em países como Alemanha, Países Baixos e Áustria.

Características Comuns:

- Urbanização intermediária.
- PIB médio equilibrado *intra-cluster*.
- Regiões com economia em desenvolvimento estável e menos dependência agrícola.

Exemplos de Regiões:

- Alemanha: Regiões industriais intermediárias (Nordrhein-Westfalen).
- Países Baixos: Flevolândia (NL33).
- Itália: Regiões do centro e noroeste.

Capitais Presentes:

"BE10" (Bruxelas), "ITI4" (Roma), "CY00" (Nicósia), "LU00" (Luxemburgo), "NL32" (Utrecht), "AT13" (Viena).

Lista de regiões abrangidas:

"BE10", "BE21", "BE22", "BE23", "BE25", "BE32", "BE33", "BE34", "BE35", "DK02", "DK03", "DK04", "DK05", "DE22", "DE23", "DE24", "DE26", "DE40", "DE50", "DE60", "DE72", "DE73", "DE80", "DE91", "DE92", "DE93", "DE94", "DEA3", "DEB1", "DEB2", "DEB3", "DEC0", "DED2", "DED4", "DED5", "DEE0", "DEF0", "DEG0", "IE04", "IE05", "EL52", "ES11", "ES12", "ES13", "ES21", "ES22", "ES23", "ES24", "ES41", "ES42", "ES43", "ES51", "ES52", "ES53", "ES61", "ES62", "ES70", "FRB0", "FRC1", "FRC2", "FRD1", "FRD2", "FRE1", "FRE2", "FRF1", "FRF2", "FRF3", "FRG0", "FRH0", "FRI1", "FRI2", "FRI3", "FRJ1", "FRJ2", "FRK1", "FRK2", "FRL0", "FRM0", "FRY1", "FRY2", "FRY4", "ITC2", "ITF3", "ITF4", "ITF6", "ITG1", "ITG2", "ITH1", "ITH2", "ITI2", "ITI4", "CY00", "LU00", "NL11", "NL12", "NL13", "NL21", "NL22", "NL23", "NL31", "NL32", "NL33", "NL34", "NL41", "NL42", "AT11", "AT12", "AT13", "AT21", "AT22", "AT31", "AT32", "AT33", "AT34", "PT15", "PT18", "PT20", "PT30", "FI19", "FI1C", "FI1D", "FI20", "SE12", "SE21", "SE22", "SE23", "SE31", "SE32", "SE33"

Cluster 3: Emergência Oriental e de Portugal

Descrição:

Esse Cluster agrupa regiões que estão em transição económica. Muitas dessas áreas são caracterizadas por:

- Economias emergentes baseadas na indústria leve, fabrico e agricultura.
- Urbanização moderada, com a presença de centros urbanos médios, mas longe de serem capitais de destaque.
- Desenvolvimento económico desigual, com PIB per capita inferior ao das regiões mais centrais da Europa, porém com potencial de crescimento acelerado.
- Forte dependência de setores como indústria tradicional, agricultura e, em alguns casos, turismo.

Características Comuns:

- Desenvolvimento Económico Intermediário: PIB moderado-baixo, inferior ao de outras regiões europeias. Presença de heterogeneidade dentro do cluster (entre áreas, como industriais e agrícolas, e entre regiões, países do Leste e Portugal).
- Economia Baseada tanto em Indústria como Agricultura: Fortes polos industriais na Polónia, Hungria, norte de Portugal e República Checa. Dependência da agricultura na Roménia e Bulgária.

- Urbanização Moderada: Centros urbanos de porte médio, mas longe do dinamismo das capitais mais ricas.
- Turismo em Crescimento: Regiões como Croácia Adriática e partes da República Checa e Portugal possuem grande atratividade turística.

Exemplos de Regiões:

- Letônia: Riga (LV00) lidera o crescimento económico, mas o interior tem desenvolvimento desigual.
- Norte de Portugal (PT11), segundo maior PIB do Cluster, zona de forte presença industrial.
- Estônia (EE00), Economia baseada em serviços digitais, TI e manufatura

Capitais Presentes:

“EE00” (Tallinn), “LV00” (Riga), “MT00” (Valeta).

Lista de regiões abrangidas:

“BG31”, “BG32”, “BG33”, “BG34”, “BG42”, “CZ02”, “CZ03”, “CZ04”, “CZ05”, “CZ06”, “CZ07”, “CZ08”, “EE00”, “HR02”, “HR03”, “HR06”, “LV00”, “LT02”, “HU12”, “HU21”, “HU22”, “HU23”, “HU31”, “HU32”, “HU33”, “MT00”, “PL21”, “PL22”, “PL41”, “PL42”, “PL43”, “PL51”, “PL52”, “PL61”, “PL62”, “PL63”, “PL71”, “PL72”, “PL81”, “PL82”, “PL84”, “PL92”, “PT11”, “PT16”, “RO11”, “RO12”, “RO21”, “RO22”, “RO31”, “RO41”, “RO42”, “SI03”, “SK02”, “SK03”, “SK04”.

Cluster 4: Centros Urbanos e Capitais Económicas

Descrição:

Este *cluster* agrupa as regiões mais desenvolvidas economicamente e urbanizadas, com média de urbanização mais alta e PIB mais elevado. Inclui as capitais e grandes centros económicos, industriais e financeiros da Europa, com infraestrutura avançada e economias diversificadas.

Características Comuns:

- Alta urbanização.
- Alto PIB.
- Grandes centros industriais, metropolitanos e capitais nacionais.

Exemplos de Regiões:

- Alemanha: DE30 (Berlim), DE50 (Estugarda).
- França: FR10 (Paris).
- Irlanda: IE06 (Dublin).
- Espanha: ES30 (Madrid).
- Portugal: PT17 (Lisboa).

Capitais Presentes:

“BG41” (Sofia), “CZ01” (Praga), “DK01” (Copenhague), “DE30” (Berlim), “IE06” (Dublin), “EL30” (Atenas), “ES30” (Madrid), “FR10” (Paris), “LT01” (Vilnius), “HU11” (Budapeste), “PL91” (Varsóvia), “PT17” (Lisboa), “RO32” (Bucareste), “SI04” (Liubliana), “SK01” (Bratislava), “FI1B” (Helsínquia).

Lista de regiões abrangidas:

"BE24", "BE31", "BG41", "CZ01", "DK01", "DE30", "IE06", "EL30", "ES30", "FR10", "HR05", "LT01", "HU11", "PL91", "PT17", "RO32", "SI04", "SK01", "FI1B", "SE11".

Conclusão

As análises feitas neste estudo permitiram-nos identificar padrões regionais significativos entre as regiões NUTS 2 da União Europeia, com base em indicadores socioeconómicos relacionados à empregabilidade. Através da Análise de Componentes Principais (PCA) foi possível reduzir a dimensionalidade dos dados, permitindo uma representação mais eficiente e compacta dos mesmos, e ainda assim manter 86% da variância total em apenas cinco componentes principais. Estas componentes demonstram dimensões críticas como emprego geral, emprego em setores especializados, e desafios estruturais do mercado de trabalho.

A análise de *Clustering*, reuniu as regiões estudadas em quatro grupos principais, baseados em características distintas de cada um:

- *Cluster 1: (Trabalho Árduo)* - Regiões caracterizadas por altos níveis de horas trabalhadas, com grande diversidade económica e ausência de capitais nacionais.
- *Cluster 2: (Europa Tradicional)* - Regiões intermediárias em termos de PIB e urbanização, refletindo estabilidade económica e diversificação de setores.
- *Cluster 3: (Emergência Oriental e de Portugal)*: Regiões com desafios socioeconómicos, PIBs mais baixos e grande dependência de setores agrícolas e industriais pouco produtivos.
- *Cluster 4: (Centros Urbanos e Capitais Econômicas)*: Grandes centros urbanos com alta urbanização e PIB elevado, refletindo o dinamismo económico da Europa.

A identificação de padrões regionais pode ajudar na criação de políticas públicas mais eficazes, auxiliando a alocação de fundos europeus, como os de coesão, de forma mais direcionada. Além disso, os resultados podem embasar estratégias de desenvolvimento regional, promovendo maior equidade e eficiência na gestão de recursos. Por exemplo, regiões do Cluster 3 poderiam ser priorizadas em iniciativas de suporte à transição económica, enquanto o Cluster 4 poderia receber incentivos para fomentar inovação e competitividade global. Os achados desse estudo revelam a importância de uma abordagem baseada em dados para enfrentar os desafios socioeconómicos atuais, ao contribuir para auxiliar os tomadores de decisão a agir de maneira informada e estratégica em prol do desenvolvimento sustentável da União Europeia.

Bibliografia

European Commission. (n.d.). *NUTS Maps*. Eurostat. Disponível em <https://ec.europa.eu/eurostat/web/nuts/maps>

European Commission. (2022). *Unemployment statistics at regional level*. Eurostat. Disponível em: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Unemployment_statistics_atRegional_level

European Commission. (n.d.). *Territorial typologies manual - degree of urbanisation*. Eurostat. Disponível em https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Territorial_typologies_manual_-_degree_of_urbanisation

Ministry for Foreign Affairs of Finland. (n.d.). *The special status of the Åland Islands*. Disponível em <https://um.fi/the-special-status-of-the-aland-islands>

European Commission. (2019) *Brexit: UK leaves EU after 47 years*. Disponível em <https://ec.europa.eu/newsroom/councilieu/items/663312/en>

Van Buuren, S. (2023). *mice: Multivariate imputation by chained equations (Version 3.17.0)*. R Documentation. Consultado a 9 de Dezembro de 2024. Disponível em: <https://www.rdocumentation.org/packages/mice/versions/3.17.0/topics/mice>

European Commission. (n.d.). *Labour Market Information:Spain*. Disponível em: https://eures.europa.eu/living-and-working/labour-market-information-europe/labour-market-information-spain_en

Money Week(2024).*Is Milan becoming Europe's new financial hub?*Disponível em: <https://moneyweek.com/economy/eu-economy/is-milan-becoming-europes-new-financial-hub>

European Comission(2022) *Regions in Europe:2022 Interactive Edition*. Disponível em: <https://ec.europa.eu/eurostat/cache/digpub/regions/#total-population>

Anexo A

Preparação dos Dados

Recolha e Tratamento

Tendo em conta o tema da empregabilidade, realizámos a exportação de dezenas de *datasets* relacionados com este tema provenientes do *website* da Organização Estatística da União Europeia (Eurostat). Um dos critérios de seleção destes *datasets* foi estarem organizados por NUTS 2 e, por isso, só foi realizada exportação de *datasets* com este critério.

Após a recolha dos *datasets*, estes foram importados para o *software R* totalizando 59 ficheiros em formato tsv.

Identificação das Regiões NUTS2

De forma a ser possível realizar a análise de acordo com os NUTS 2, foi criado um documento auxiliar que continha os códigos e nome de cada região ao nível dos NUTS 2 dos países da União Europeia.

Este ficheiro foi alterado algumas vezes ao longo deste processo pelas seguintes razões:

1. No ano escolhido para a análise, o Reino Unido ainda fazia parte da UE e, por isso, considerámos mantê-lo na lista de NUTS2. No entanto, devido à sua saída em 2020 e à atualização dos *datasets* por parte da EuroStat ao longo dos anos, muitos destes não têm qualquer registo relativo ao Reino Unido. Isto fez com que a quantidade de valores omissos nos *datasets* fosse muito superior. Por esse motivo, devido à inexistência de dados do Reino Unido causada pela atualização dos *datasets* da Eurostat, decidimos retirar todas as suas regiões da nossa lista, ficando com 269.
2. Surgiu mais tarde outra questão relativamente às regiões chamadas "Extra-Regio". Estas são regiões adicionais que não correspondem a regiões geográficas típicas e que por isso não têm cálculo de população nem de PIB per capita. Cada país tem uma destas regiões.

Nestas regiões estão incluídos dados que não podem ser atribuídos às outras regiões NUTS de um país, tais como: atividades em zonas marítimas, embaixadas, bases militares no estrangeiro, consulados e atividades transnacionais. A maior parte destas "Extra-Regio" não apresenta nenhum tipo de dados nos nossos *datasets* e seria incoerente mantê-las para alguns países e não

para outros, por isso, decidimos retirá-las também da lista. Ficamos então com uma lista final de **242** regiões ao nível NUTS2 dos países da UE.

Limpeza inicial dos datasets

Demos início à limpeza de dados dos *datasets*, com a uniformização da coluna que continha o código de cada região. Esta coluna continha informação irrelevante para a análise, por isso, ao extrair os códigos das regiões e criar uma nova coluna com estes, garantimos uma visualização fácil das regiões associadas a estes dados.

Foram realizadas outras alterações a cada *dataset* como: remoção de linhas em que o código não correspondesse a nenhum código dos NUTS 2; remoção das colunas 2010 e 2011, pois a crise financeira aumentava a probabilidade dos dados estarem distorcidos, o que poderia afetar a imputação; e remoção das colunas com o nome das regiões e código do país.

Análise Exploratória

Valores Omissos e Escolha do Ano a ser Analisado

De forma a avaliar a qualidade dos dados no que diz respeito a valores omissos, foi criada uma função de cálculo que retorna a média dos valores omissos em todos os *datasets*, segmentada por ano (entre 2012 e 2023).

Entre 2013 e 2019, os valores omissos variaram entre 5,59% e 6,90%, enquanto nos anos seguintes, a quantidade de valores omissos foi superior. Considerando a pequena variação nos valores omissos entre os anos de 2013 e 2019, optou-se por escolher o ano de 2019 para realizar uma análise mais completa e atualizada possível.

Após a seleção do ano sobre o qual incidiria a análise, procedeu-se à criação do novo *dataset* que agritará todos os dados do ano de 2019 de cada variável e será o nosso objeto de análise. Para tal, extraiu-se a coluna correspondente ao ano de 2019 de cada *dataset*, sendo que as linhas de cada coluna foram verificadas com o código NUTS 2, para garantir que não existissem desalinhamentos. Em cada iteração, as colunas foram renomeadas de acordo com o nome do respetivo *dataset*, facilitando a análise de cada variável. O resultado final foi uma base de dados com 60 colunas (das quais 59 variáveis) e 242 linhas.

Após a criação da base de dados, foi necessária uma verificação detalhada da quantidade de valores omissos ("." e "NA") em cada variável, com o objetivo de identificar variáveis que poderiam ser eliminadas por falta de informação. Para este processo, foi implementada uma função de contagem, que identificou as variáveis com mais de 20% de

valores omissos, tendo estas sido removidas. Como resultado, foram eliminadas 8 variáveis, ficando o *dataset* com 51 variáveis INPUT.

Estudo da Normalidade das Variáveis

A seguir, para estudar a normalidade das variáveis selecionadas, foi realizado o teste de Shapiro-Wilk, complementado com a análise dos gráficos Quantil-Quantil (Q-Q). Na Figura 5, é possível observar um exemplo dos gráficos Q-Q, abrangendo as variáveis “CompEmp_Pub”, “Cult_Emp”, “HR_Tech” e “RL_prod”.

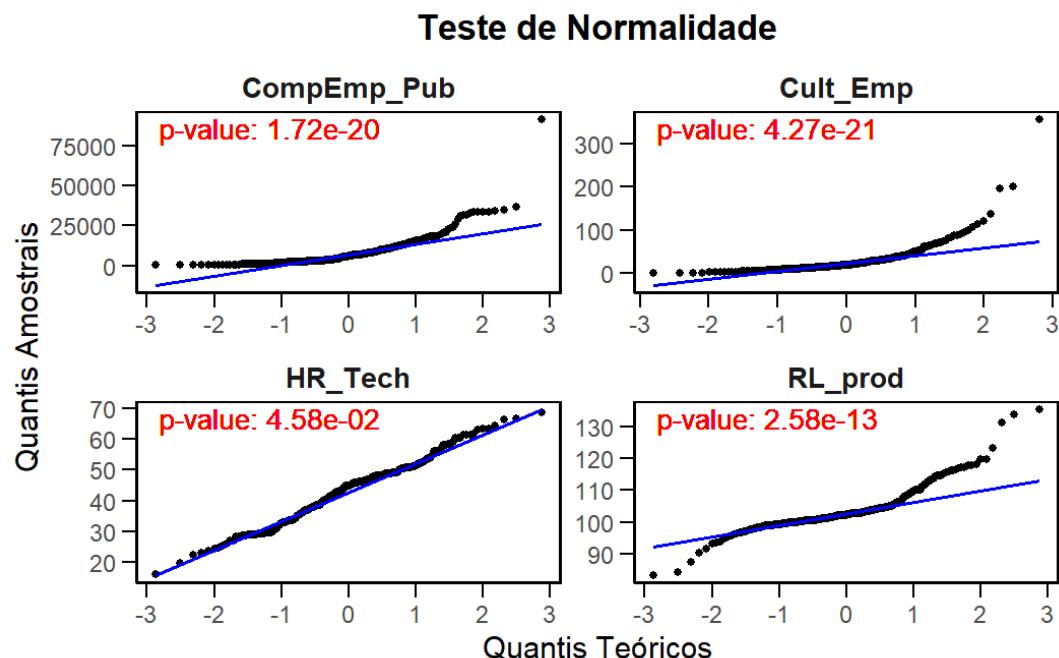


Figura 5 - Teste de Normalidade para as variáveis “CompEmp_Pub”, “Cult_Emp”, “HR_Tech” e “RL_prod”.

Os gráficos Q-Q revelaram desvios significativos da normalidade para a maioria das variáveis analisadas, similar às variáveis “CompEmp_Pub”, “Cult_Emp” e “RL_prod”, que apresentaram desvios notáveis nos extremos da distribuição, enquanto “HR_Tech” exibiu um comportamento mais próximo da normalidade. Os resultados do *p-value* corroboraram estas observações visuais, com as variáveis “CompEmp_Pub”, “Cult_Emp” e “RL_prod” a registarem *p-values* muito baixos, indicando forte evidência contra a hipótese de normalidade. A variável “HR_Tech”, por sua vez, obteve um *p-value* de 0,045, ligeiramente abaixo do limiar de 0,05, mas ainda insuficiente para ser considerada normalmente distribuída.

Análise de Regiões Ausentes

No que toca a valores omissos, quando não existem dados disponíveis no Eurostat, os campos são representados pelo indicador “:”. Antes de proceder à imputação destes valores,

foi necessário avaliar outro tipo de dados omissos, representados como “NA”. Os campos assinalados como “NA” resultam da ausência da região correspondente no *dataset* original. Após a análise, verificou-se a existência de 52 campos com valores “NA”, distribuídos por 6 variáveis.

Ao estudar em detalhe a base de dados, observámos os seguintes casos de dados omissos (NA):

- Nas variáveis “NEET_M” e “NEET_F”, a região de Mayotte (FRY5) está ausente.
- Nas variáveis “Y_EmpRate_M”, “Y_EmpRate_F” e “Y_EmpRate_Age”, existem três campos com “NA” em cada uma. Contudo, o valor está presente sob o código de região antigo, HR04, que em 2021 foi subdividido em três novas regiões: HR02, HR05 e HR06. Assim, os valores correspondentes existem, estando apenas associados à região antiga, sem a respetiva subdivisão.
- Na variável “Cult_Emp”, foram encontrados 42 campos com “NA”, que correspondem a todas as regiões de Espanha, todas as regiões da Polónia e algumas regiões de territórios extra-marítimos de França.

Como se pode observar na Tabela 1, a quantidade de valores omissos é relativamente reduzida em cada variável, sendo que as variáveis não apresentadas na tabela não contêm quaisquer valores omissos.

Variável	Percent_NA(%)	Percent_":"(%)
LT_Unemp_Age	0.00	9.92
Emp_HTech	0.00	8.68
Emp_Tech_Total	0.00	8.68
AWH_Total_65p	0.00	6.20
Emp_65p	0.00	4.96
EA_Total_65p	0.00	4.96
NEET_M	0.41	2.48
Emp_PT_F_15_64	0.00	2.48
Emp_PT_M_15_64	0.00	2.07
Cult_Emp	17.36	1.65
NEET_F	0.41	1.24
Unemp_CZ	0.00	0.83
Y_EmpRate_M	1.24	0.41
Y_EmpRate_F	1.24	0.41
UnempRate	0.00	0.41
PopUnemp_15p	0.00	0.41
Emp_15_64_LE	0.00	0.41
Y_EmpRate_Age	1.24	0.00

Tabela 1: Percentagem de valores omissos presentes nas variáveis escolhidas

Quanto às variáveis que continham campos com o indicativo “：“, foi analisado se estas tinham uma maior frequência em alguma região em específico. Apenas se destacaram as regiões FI20 (Ilhas Åland, Finlândia), com 17 valores omissos, e PT20 (Açores), com 8 valores omissos. As restantes regiões com valores omissos apresentam esses valores em cinco variáveis ou menos.

Após alguma investigação, pode-se deduzir que os valores omissos na região FI20 se devem possivelmente à baixa densidade populacional e características únicas da região. Estas observações podem justificar o indicativo “u” (*low reliability*) presente nestes valores, uma vez que estes continuam consistentemente ausentes não só em 2019, mas também nos anos anteriores. Segundo a mesma lógica, também a região PT20, por ser uma ilha predominantemente rural, oferece uma explicação plausível para a falta de dados relacionados com setores tecnológicos.

Imputação de Valores Omissos e Análise de Outliers

Antes de iniciar esta secção, é importante salientar que a maioria das variáveis analisadas não precisou de imputação ou apresentou apenas um número reduzido de valores omissos a serem imputados. Contudo, algumas variáveis apresentaram um número de valores omissos que consideramos significativo, o que exigiu um estudo detalhado da distribuição e dispersão dos dados imputados, de forma a assegurar a sua qualidade. Para ilustrar ambos os casos, foram selecionadas as variáveis “Y_EmpRate_F” e “Cult_Emp”.

Outliers

Antes de proceder à imputação dos valores omissos, foi realizada uma análise dos *outliers*, uma vez que estes podem ter um impacto significativo no processo de imputação. Para uma visualização mais clara dos dados, foram utilizados diagramas de caixa (*boxplots*), estando ambos diagramas para “Y_EmpRate_F” e “Cult_Emp” presentes nas Figuras 6 e 7.

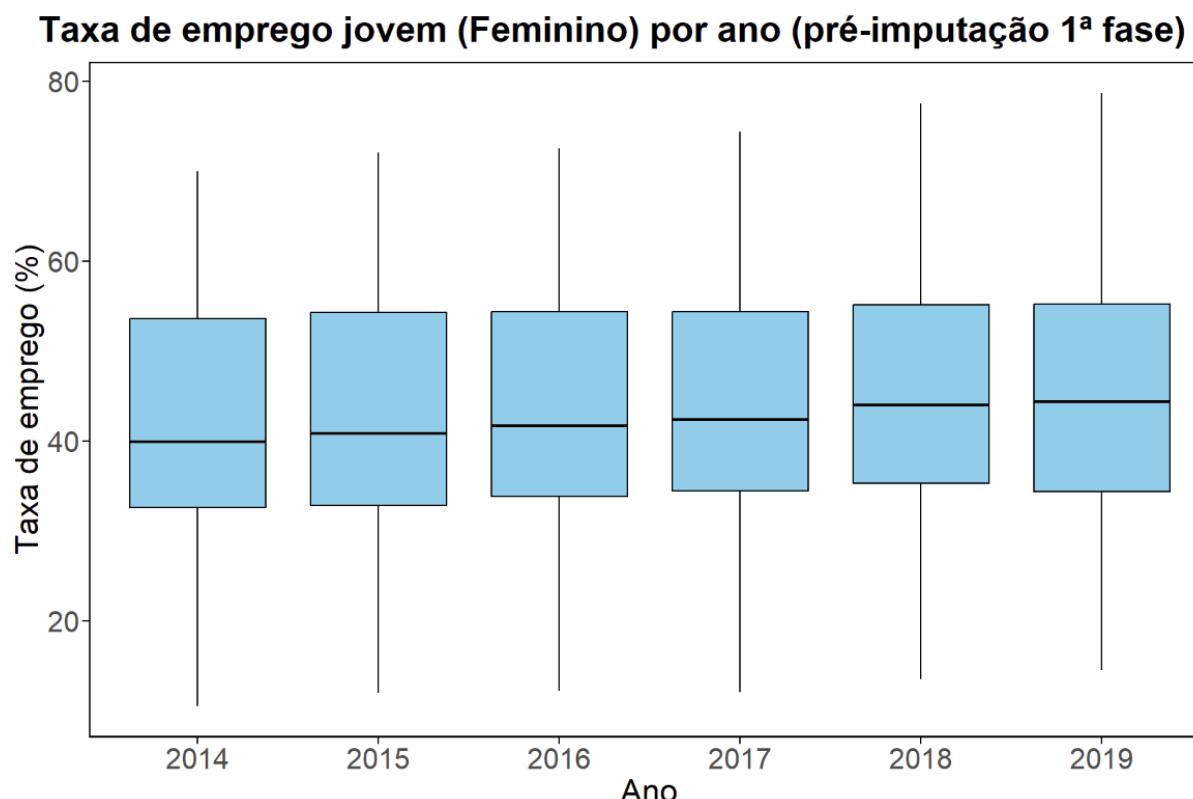


Figura 6 - Diagrama de caixas para análise de outliers da variável ”Y_EmpRate_F“, antes da 1ª fase de Imputação, para os anos 2014 a 2019

Emprego no setor cultural por ano (pré-imputação 1ª fase)

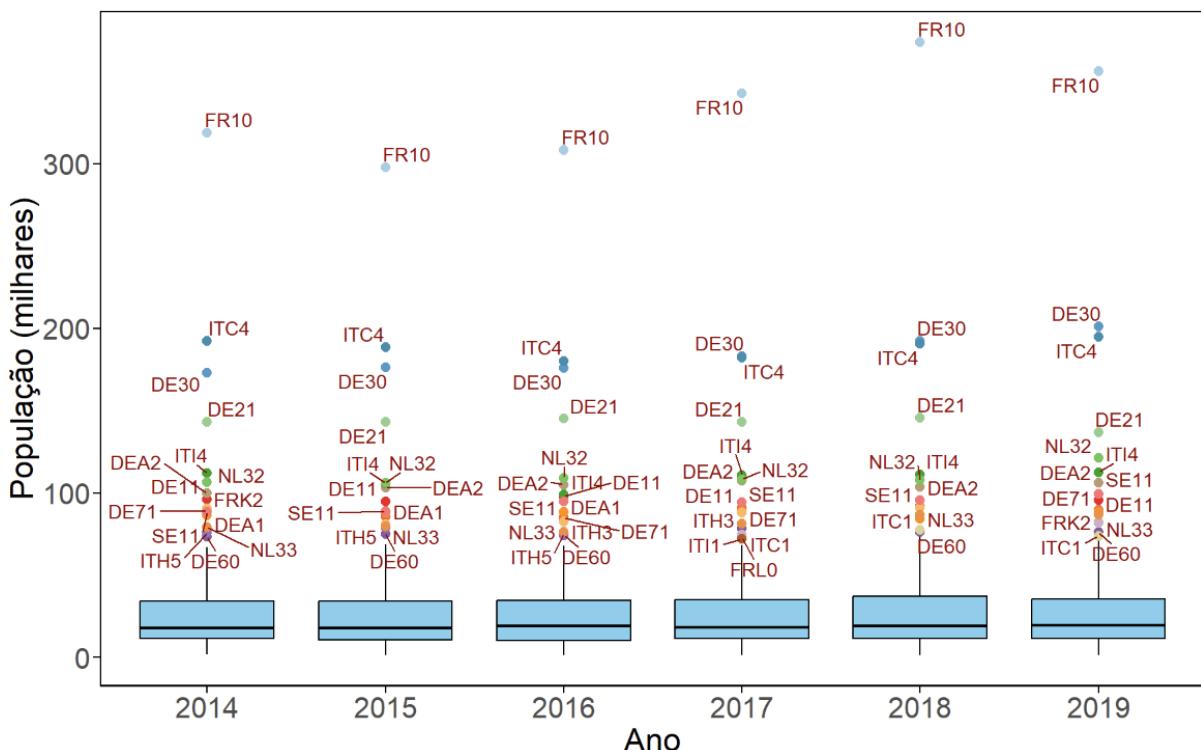


Figura 7 - Diagrama de caixas para análise de outliers da variável "Cult_Emp", antes da 1ª fase de Imputação, para os anos 2014 a 2019

O diagrama referente à variável "Y_EmpRate_M" não apresenta valores atípicos, com todos os dados entre 2014 e 2019 situados dentro dos limites inferior e superior dos quartis. Em contrapartida, o diagrama da variável "Cult_Emp" revela a presença de regiões que se destacam como *outliers* em vários anos. Este padrão foi observado em diversas variáveis, especialmente naquelas associadas a níveis superiores de empregabilidade ou maior compensação monetária.

Algumas dessas regiões, que estão presentes como *outliers* em 25 ou mais variáveis, incluem: ITC4 (Milão), FR10 (Paris), ES51 (Barcelona), DE21 (Munique), ITI4 (Roma) e ES30 (Madrid). Esta amostra fornece um indicativo relevante da relação entre regiões que são grandes polos económicos, culturais e turísticos, e a sua frequência como *outliers* em várias variáveis. As regiões que são *outliers* frequentes estão presentes na Tabela 2.

Outlier_Reg	ITC4	FR10	FRK2	ES51	DE21	ITI4	ES61	DE11	DEA1	ES30
Frequência	35	32	30	29	28	27	26	25	25	24

Tabela 2 – As dez regiões com maior frequência de ocorrência como outliers nas variáveis analisadas

No entanto, para outras regiões, não existia uma explicação tão clara para a sua presença enquanto *outliers*. Por este motivo, decidimos verificar se este enquadramento se devia a dados com uma baixa confiabilidade através da análise da presença dos indicativos (*flags*) fornecidos no site do Eurostat.

O tipo de indicativos ou *flags* que podem ser atribuídos aos dados são:

- (p) *provisional*
- (e) *estimated*
- (d) *definition differs*
- (u) *low reliability*
- (b) *break in time series*
- (bu) *break in time series, low reliability*
- (bd) *break in time series, definition differs*
- (bdu) *break in time series, definition differs, low reliability*
- (du) *definition differs, low reliability*

Foi possível verificar que os dados caracterizados como *outliers* apresentavam poucas *flags*, com apenas uma ou duas regiões entre os *outliers*. A *flag* mais frequente foi a "(b)" (*break in time series*), que indica a existência de uma descontinuidade ou mudança nos dados em determinado ano, o que pode afetar a comparação feita com outros anos. Esta *flag* esteve predominantemente presente nas regiões da Polónia e, após uma pesquisa, foi possível verificar que todas as regiões da Polónia apresentavam esta *flag* no ano de 2019 e em diversos *datasets*. Isto sugere que em 2019 ocorreu algum evento atípico relacionado com a recolha ou análise dos dados neste país. Esta análise confirma também que, para este caso em específico, a *flag* "(b)" nos *outliers* não indica erros nos dados, sendo assim um dado de confiança.

Uma situação semelhante ocorreu com as regiões da Alemanha, em que constavam as *flags* "(e)" (*estimated*) e "(p)" (*provisional*) nas variáveis relacionadas com compensação monetária de funcionários.

Não é possível tirar uma conclusão sobre a razão que leva à atribuição deste indicativo. Uma hipótese é que o início da pandemia tenha causado uma reavaliação dos dados ou um atraso no seu fornecimento. Alternativamente, poderá também ter sido uma situação arbitrária.

No extremo oposto dos *outliers*, foi possível observar que, por exemplo, no que diz respeito à taxa de empregabilidade, algumas regiões surgem como *outliers* no limite inferior, como é o caso de: FRY4, FRY3, FRY5, ITG1, ITF3, ITF6 e ITF4. Estas regiões incluem territórios extra-marítimos de França e regiões do Sul de Itália. Regiões estas que historicamente apresentam um desempenho económico mais fraco.

Esta análise permitiu compreender melhor os dados, identificando as regiões que se destacam nas variáveis analisadas, dando também um bom indicativo de regiões que

poderão agrupar-se no mesmo *cluster*. No entanto, apenas com a análise de PCA e *Clustering* será possível obter conclusões mais robustas e confiáveis.

Imputação:MICE e Métodos

Tendo decidido manter os *outliers* neste estudo, dada a sua importância para esta análise, procedeu-se para a fase de imputação de dados. Para as imputações, recorremos ao pacote MICE (*Multivariate Imputation by Chained Equations*; Van Buuren, 2023), que é amplamente utilizado para lidar com dados ausentes em análises multivariadas. Este pacote permite imputar valores ausentes através de métodos iterativos que consideram as relações entre as variáveis, gerando múltiplos conjuntos de dados imputados.

Existem diversos métodos para a imputação, dependendo das características dos dados. Entre os diversos métodos disponíveis encontram-se o método "norm", que pressupõe que as variáveis seguem uma distribuição normal. No entanto, os gráficos Q-Q apresentados na Figura 5 invalidam esta hipótese, inviabilizando a aplicação deste método. Outra alternativa é o método CART, que utiliza Árvores de Decisão para prever os valores mais adequados com base em padrões nos dados. Contudo, optámos por outras abordagens que considerámos mais adequadas às especificidades dos dados analisados. De modo a alcançar a melhor qualidade de dados possível, as imputações foram realizadas em duas fases:

- Primeira fase: Imputações realizadas em cada *dataset* individualmente para o ano de 2019, utilizando como preditores todos os anos anteriores disponíveis, excluindo os anos 2011 e 2012. Foi utilizado o método *Predictive Mean Matching* (PMM), que é particularmente eficaz para lidar com dados numéricos e se adapta bem a séries temporais curtas.
- Segunda fase: Imputações realizadas no *dataset* final, obtido após a primeira fase de imputação, e que continha o ano escolhido para todas as variáveis. O método selecionado foi o *Random Forest* (RF), devido à sua capacidade de captar relações complexas entre variáveis, robustez em situações de dados ausentes e utilização da informação disponível em outras variáveis para realizar previsões.

Imputação: Primeira Fase

Na primeira fase de imputações, foram capturados os Padrões de Dados Ausentes, representados nas Figuras 8 e 9, de modo a fornecer uma representação mais visual das imputações realizadas na coluna correspondente aos dados de 2019 (coluna X2019).

Padrão de Dados Ausentes - Y_EmpRate_F (pré e pós-imputação 1^a fase)

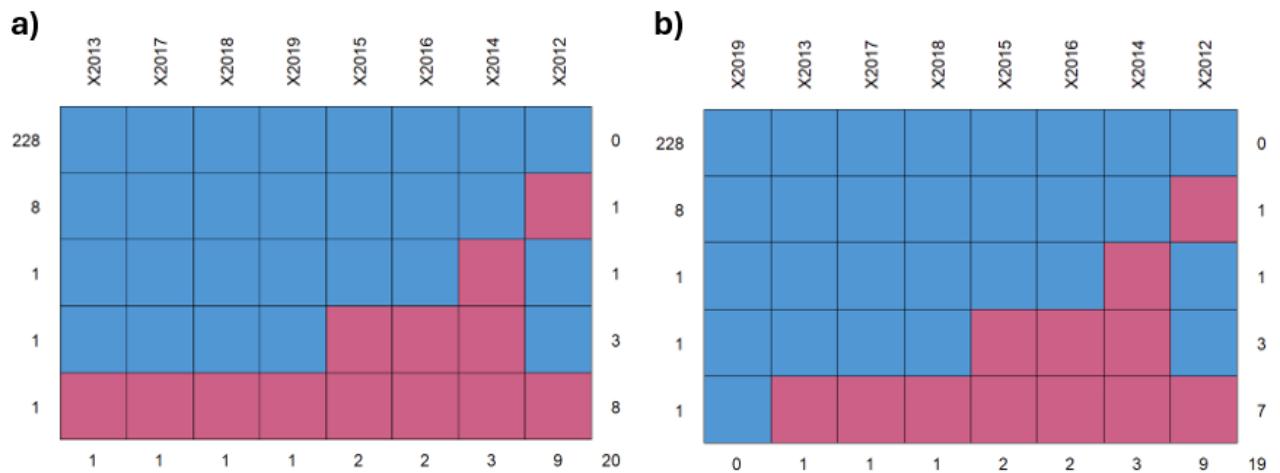


Figura 8 - Padrão de Dados Ausentes para a variável “Y_EmpRate_F”: (a) antes e (b) após a 1^a fase de Imputação

Padrão de Dados Ausentes - Cult_Emp (pré e pós-imputação 1^a fase)

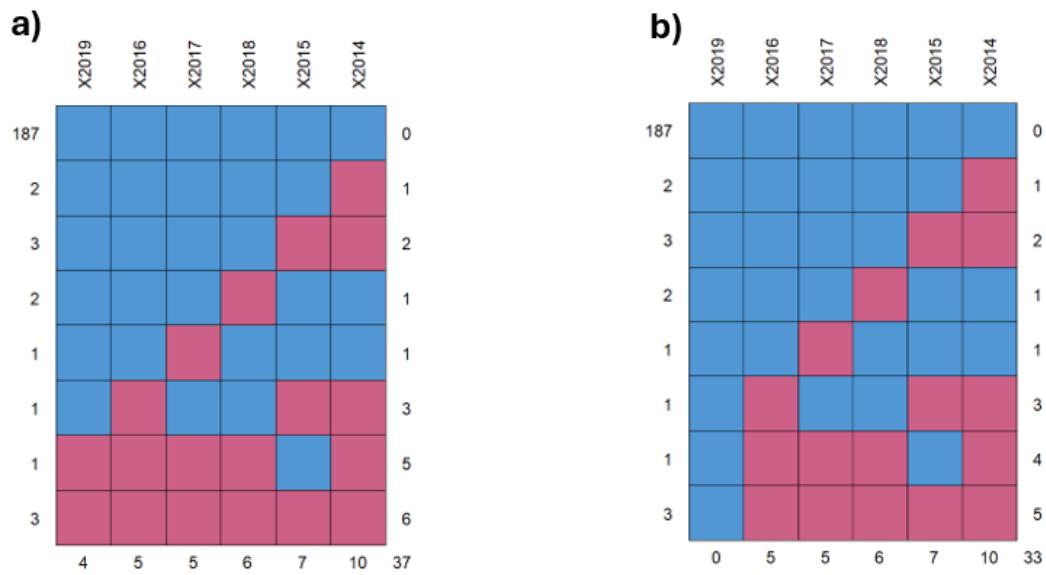


Figura 9 - Padrão de Dados Ausentes para a variável “Cult_Emp”: (a) antes e (b) após a 1^a fase de Imputação

Para interpretar os resultados apresentados nos padrões, cada linha no diagrama representa uma observação (neste caso, uma região NUTS), enquanto cada coluna corresponde a uma variável ou a um ano específico. As células indicam a presença (azul-claro) ou a ausência (cor-de-rosa) de dados para cada combinação de observação e variável/ano.

Para a variável “Y_EmpRate_F”, observa-se que apenas existia um valor ausente para o ano 2019, localizado numa linha onde também estavam ausentes os valores dos anos anteriores. Embora o valor tenha sido imputado com sucesso, a ausência de preditores

específicos para a mesma região nos outros anos justifica um estudo mais aprofundado do valor imputado, pois, apesar de ser apenas um valor ausente, a sua imputação pode ter influenciado a distribuição dos dados.

Uma situação semelhante foi observada nos padrões da variável “Cult_Emp”. Para um dos valores ausentes no ano de 2019, os dados também estavam ausentes nos anos anteriores, com exceção de 2015. Quanto aos três valores ausentes restantes, verificou-se a ausência de dados para todos os anos anteriores. A imputação para o ano de 2019 foi bem-sucedida, com um total de quatro valores imputados.

Qualidade da Imputação

Para avaliar a qualidade das imputações realizadas, foi utilizado um gráfico de dispersão, no qual ambas as variáveis para o ano de 2019 estão representadas na Figura 10.

Estudo da Qualidade das Imputações - Y_EmpRate_F e Cult_Emp (1^a fase)

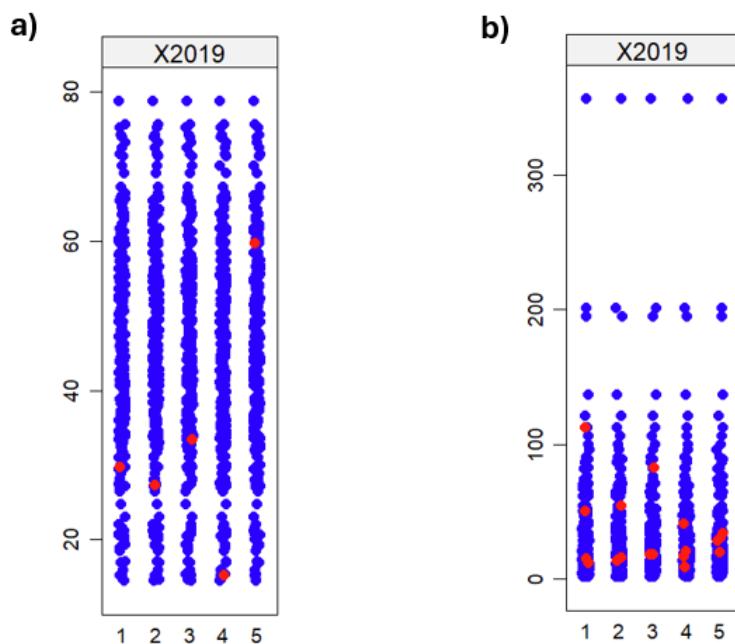


Figura 10 – Qualidade das imputações realizadas para o ano 2019, para as variáveis: a) “Y_EmpRate_F” e b) “Cult_Emp”, após a 1^a fase de Imputação

Nos gráficos de dispersão, observou-se que os pontos vermelhos, que representam os valores imputados, estavam sobrepostos a pontos azuis, indicando que as imputações não introduziram desvios significativos na distribuição dos dados. Este padrão sugere que as imputações mantiveram a integridade da distribuição e não afetaram visivelmente as variáveis analisadas. Adicionalmente, foram também calculadas as diferenças percentuais na variância entre os dados originais e imputados para as variáveis “Y_EmpRate_F” e “Cult_Emp” no ano de 2019, que foram de 0,07% e 0,65%, respectivamente. Estes valores indicam que as

imputações não alteraram substancialmente a variabilidade dos dados, o que reforça a qualidade do processo de imputação. A variância, neste estudo, refere-se à dispersão dos dados em torno da média, ou seja, ao grau de dispersão e à forma como os dados estão distribuídos, o que é crucial para avaliar a consistência e a integridade das imputações.

Para concluir a 1^a fase de imputações, e de forma a verificar possíveis *outliers* resultantes do processo, procedemos a estudar os diagramas de caixas, representados nas Figura 11 e 12.

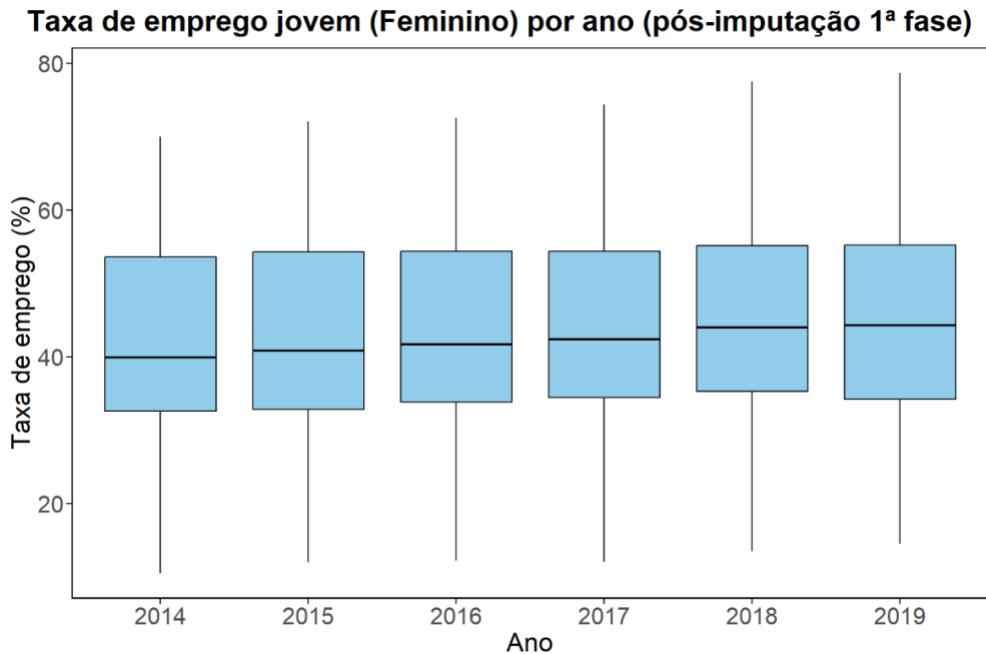


Figura 11 - Diagrama de caixas para análise de outliers da variável "Y_EmpRate_F", após a 1^a fase de Imputação, para os anos 2014 a 2019

Emprego no setor cultural por ano (pós-imputação 1^a fase)

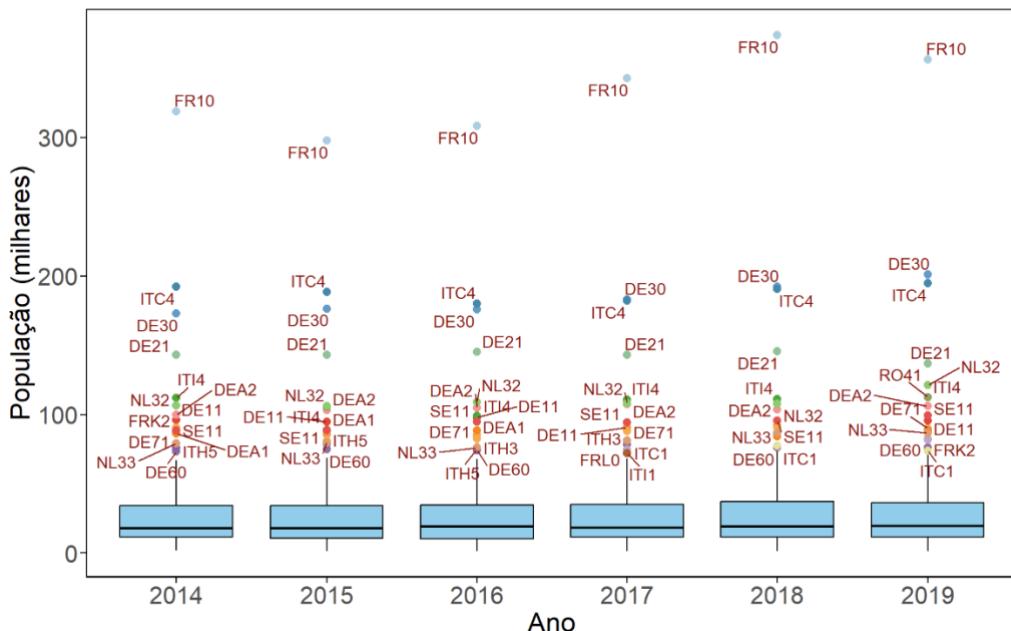


Figura 12 - Diagrama de caixas para análise de outliers da variável "Cult_Emp", após a 1^a fase de Imputação, para os anos 2014 a 2019

Para ambas as variáveis, assim como para a maioria das restantes, não foram observadas mudanças significativas em termos de outliers após a 1^a fase de imputação.

Após a conclusão da 1.^a fase, foi construído o dataset com todas as variáveis relativas ao ano de 2019. Em seguida, avançou-se para a 2.^a fase, sendo que os padrões de dados ausentes estão representados nas Figuras 13 e 14.

Padrão de Dados Ausentes para o ano 2019 (pré-imputação 2^a fase)

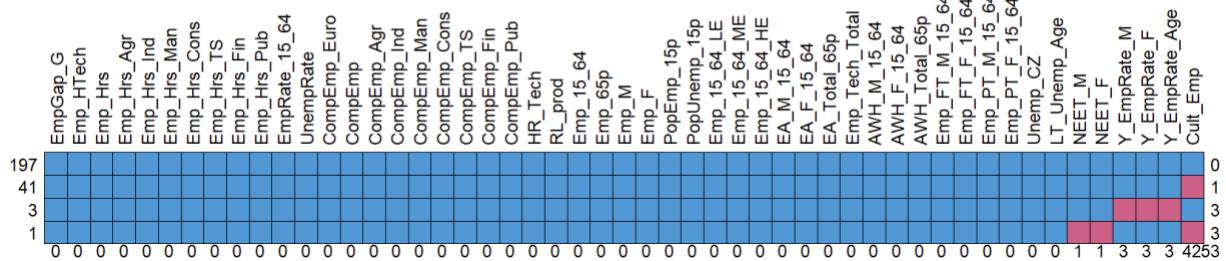


Figura 13 - Padrão de Dados Ausentes para as variáveis para o ano 2019, antes da 2^a fase de Imputação

Padrão de Dados Ausentes para o ano 2019 (pós-imputação 2^a fase)

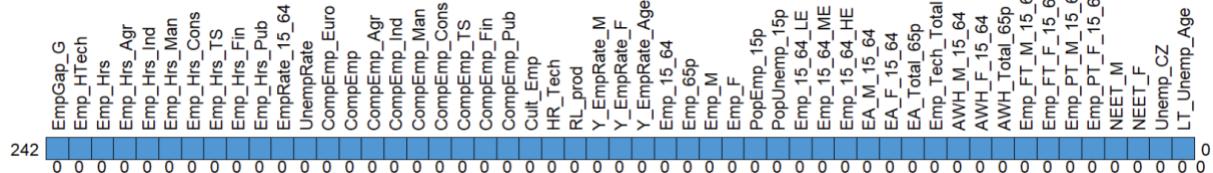


Figura 14 - Padrão de Dados Ausentes para as variáveis para o ano 2019, após a 2^a fase de Imputação

Imputação: Segunda Fase

Ao observar os padrões apresentados, conclui-se que todos os valores em falta foram imputados com sucesso após a 2.^a fase. Seis variáveis apresentaram valores ausentes a serem imputados, sendo que a variável “Cult_Emp” registou o maior número de imputações, com um total de 42 valores ausentes. Por outro lado, a variável “Y_EmpRate_F” apenas teve imputados três valores ausentes. Para avaliar a qualidade das imputações, foram gerados novos gráficos de dispersão, os quais podem ser visualizados na Figura 15.

Estudo da Qualidade das Imputações - Y_EmpRate_F e Cult_Emp (2^a fase)

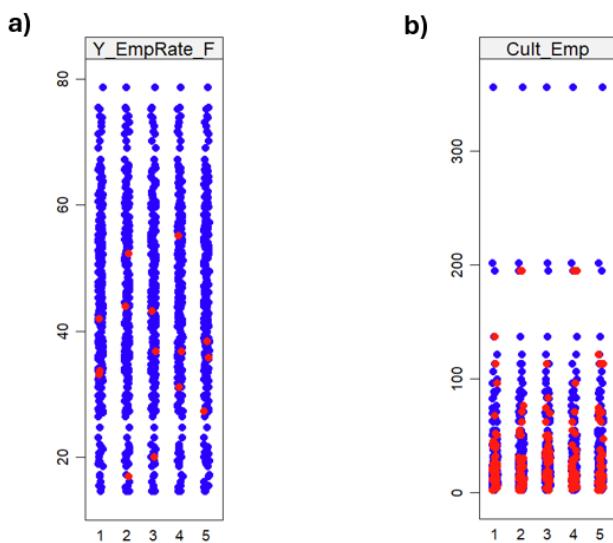


Figura 15 - Qualidade das imputações realizadas para as variáveis: a) “Y_EmpRate_F” e b) “Cult_Emp”, após a 2^a fase de Imputação

Embora as imputações para ambas variáveis pareçam seguir o padrão geral da distribuição, a quantidade substancial de valores imputados para a variável “Cult_Emp” requer uma interpretação cautelosa dos resultados. Por este motivo, e considerando a possibilidade de algum impacto na estrutura original dos dados, foram analisados tanto o aparecimento de possíveis novos *outliers*, com recurso a gráficos de caixas apresentados nas Figuras 16 e 17, como também a variação na dispersão dos dados, ilustrada no gráfico de barras na Figura 18.

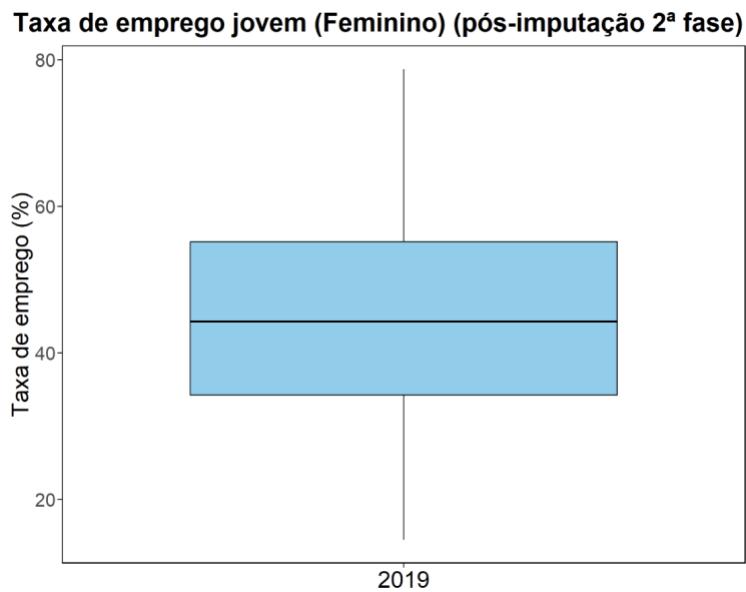


Figura 16 – Diagrama de caixas para análise de outliers da variável "Y_EmpRate_F", após a 2^a fase de Imputação, para o ano 2019

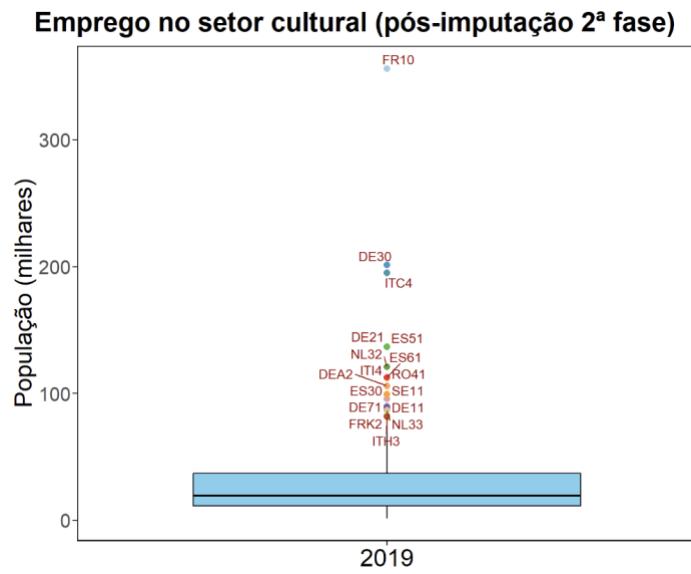


Figura 17 - Diagrama de caixas para análise de outliers da variável "Cult_Emp", após a 2^a fase de Imputação, para o ano 2019

Enquanto para a variável “Y_EmpRate_F” não foram observados novos *outliers*, na variável “Cult_Emp” verifica-se o surgimento de novas regiões, com destaque para a região ES51, que se apresenta como *outlier* em pelo menos 30 outras variáveis do estudo.

Comparação da Dispersão dos Dados (pós-imputação 2^a fase)

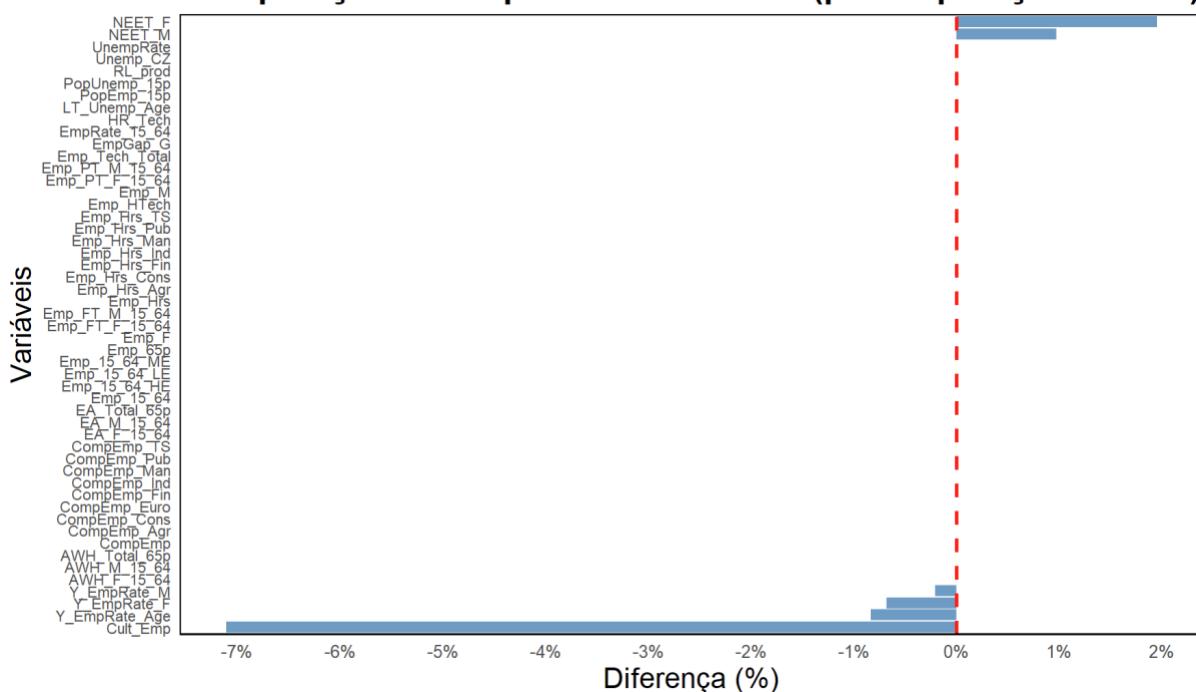


Figura 18 - Comparação da dispersão dos dados para as variáveis escolhidas, após a 2^a fase de Imputação

Ao analisar o gráfico de barras horizontal, observou-se que a variável “Cult_Emp” apresentou a maior variação na dispersão dos dados após imputação, com uma diferença de aproximadamente -7% em relação aos dados antes da imputação. Este resultado era esperado, dada a quantidade significativa de valores ausentes imputados, conforme mencionado anteriormente. No entanto, considerando que o gráfico de dispersão mostrado na Figura 15 indicou que a distribuição dos dados foi mantida, consideramos esta variação como aceitável, e mantivemos a variável para análises futuras.

Por outro lado, a variável “Y_EmpRate_F” teve uma variação menor, de cerca de -1%, o que também estava dentro das expectativas, uma vez que apenas três valores ausentes foram imputados. Além disso, foram observadas outras variáveis que apresentaram variações pouco significativas, com impacto mínimo na distribuição dos dados: as variáveis “NEET_F” e “NEET_M” apresentaram uma variação positiva inferior a 2%, enquanto as variáveis “Y_EmpRate_M” e “Y_EmpRate_Age” exibiram uma variação negativa inferior a 1%.

Concluída a preparação dos dados e a seleção das variáveis relevantes, procedeu-se à aplicação do PCA para identificar os componentes principais que melhor representam as variáveis em estudo.

Anexo B

PCA

Pair plot 1:



Figura 19 - Gráfico de dispersão 1

Variáveis Incluídas:

“EmpGap_G”, “Emp_HTech”, “Emp_Hrs”, “Emp_Hrs_Agr”, “Emp_Hrs_Ind”, “Emp_Hrs_Man”, “Emp_Hrs_Cons”, “Emp_Hrs_TS”, “Emp_Hrs_Fin”, “Emp_Hrs_Pub”, “EmpRate_15_64”, “UnempRate”, “CompEmp_Euro”.

Notas:

- As variáveis relacionadas às horas de trabalho (“Emp_Hrs_*”) apresentam correlações positivas fortes, indicando que regiões com maior volume de trabalho num setor tendem a ter volumes elevados noutros setores.
- A variável “UnempRate” mostra uma correlação negativa com “EmpRate_15_64”, como esperado. Regiões com taxas de emprego mais altas tendem a ter taxas de desemprego mais baixas.
- Existe uma relação inversa entre Emp_HTech e EmpGap_G. Regiões mais avançadas tecnologicamente parecem ter lacunas menores no emprego.
- Na maior parte das variáveis observam-se os *outliers*.

Pair plot 2:

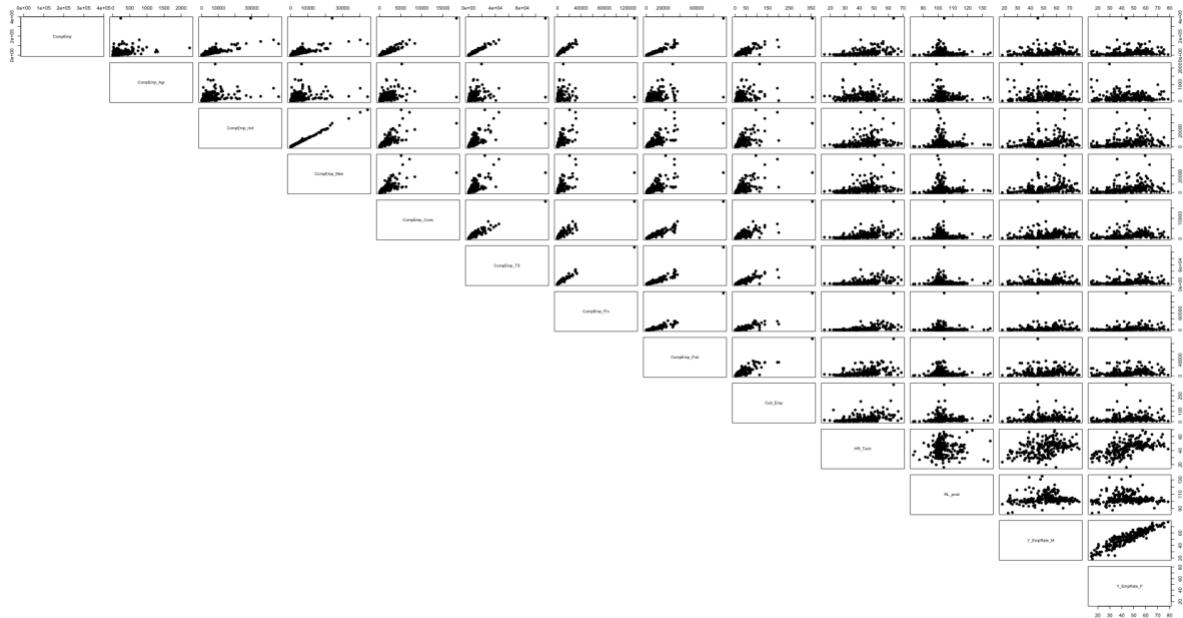


Figura 20 - Gráfico de dispersão 2

Variáveis Incluídas:

“CompEmp”, “CompEmp_Agr”, “CompEmp_Ind”, “CompEmp_Man”,
 “CompEmp_Cons”, “CompEmp_TS”, “CompEmp_Fin”, “CompEmp_Pub”, “Cult_Emp”,
 “HR_Tech”, “RL_prod”, “Y_EmpRate_M”.

Notas:

- Existe uma correlação linear muito forte entre as variáveis de compensação (“CompEmp_*”). Por exemplo, regiões com altos valores de compensação no setor de indústria (“CompEmp_Ind”) também apresentam altos valores no setor de fabrico (“CompEmp_Man”).
- A compensação no setor público (“CompEmp_Pub”) apresenta menor correlação com os setores privados, refletindo diferenças estruturais.
- O emprego cultural (“Cult_Emp”) não apresenta correlações fortes com outras variáveis, sugerindo que este tipo de emprego pode ser influenciado por fatores regionais únicos.

Pair plot 3:

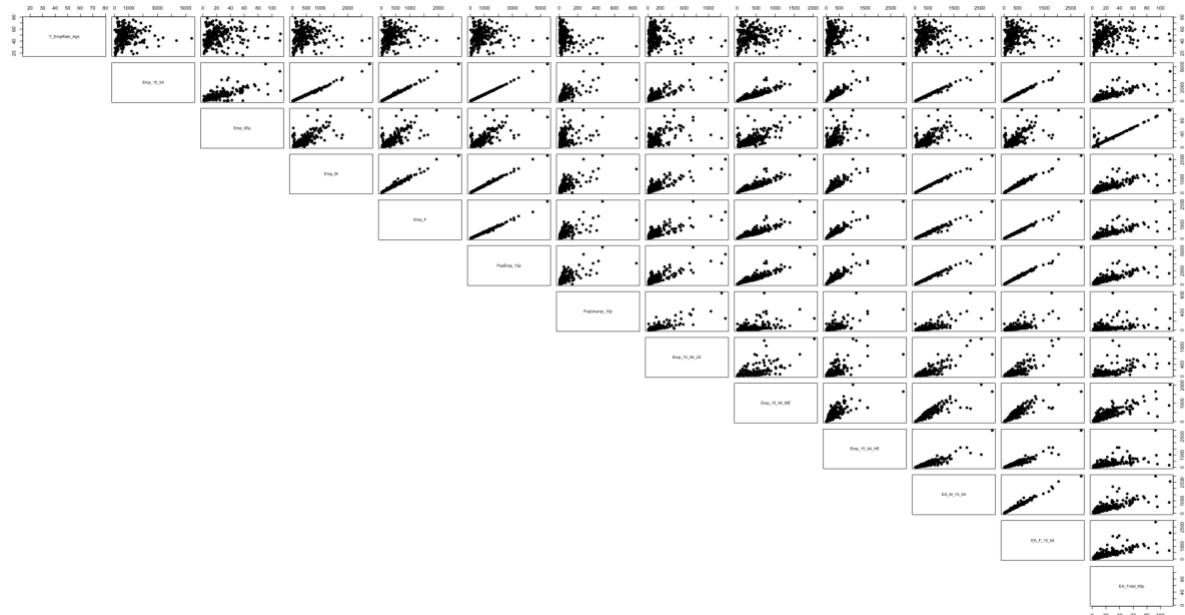


Figura 21 – Gráfico de dispersão 3

Variáveis Incluídas:

“Y_EmpRate_Age”, “Emp_15_64”, “Emp_65p”, “Emp_M”, “Emp_F”, “PopEmp_15p”,
“PopUnemp_15p”, “Emp_15_64_LE”, “Emp_15_64_ME”, “Emp_15_64_HE”, “EA_M_15_64”,
“EA_F_15_64”, “EA_Total_65p”

Notas:

- Existe uma relação positiva clara entre os níveis educacionais (“Emp_15_64_LE”, “Emp_15_64_ME”, “Emp_15_64_HE”) e as taxas de emprego. Regiões com maior proporção de trabalhadores com educação superior apresentam melhores resultados no mercado laboral.
- As taxas de emprego masculino (“EA_M_15_64”) são consistentemente mais altas do que as femininas (“EA_F_15_64”).
- As variáveis relacionadas à população empregada e desempregada mostram padrões esperados: maior população empregada está associada a menor desemprego.

Pair plot 4:



Figura 22 – Gráfico de dispersão 4

Variáveis Incluídas:

“Emp_Tech_Total”, “AWH_M_15_64”, “AWH_F_15_64”, “AWH_Total_65p”,
 “Emp_FT_M_15_64”, “Emp_FT_F_15_64”, “Emp_PT_M_15_64”, “Emp_PT_F_15_64”,
 “NEET_M”, “NEET_F”, “Unemp_CZ”, “LT_Unemp_Age”

Notas:

- Existe uma correlação positiva entre o emprego a tempo integral masculino (“Emp_FT_M_15_64”) e feminino (“Emp_FT_F_15_64”). No entanto, o emprego parcial feminino é significativamente mais elevado do que o masculino, refletindo desigualdades estruturais no mercado laboral.
- O desemprego prolongado (“LT_Unemp_Age”) apresenta padrões distintos, sugerindo dificuldades estruturais específicas em certas regiões.
- As horas médias trabalhadas por género são similares (“AWH_M_15_64” vs “AWH_F_15_64”).

Visualização da correlação:

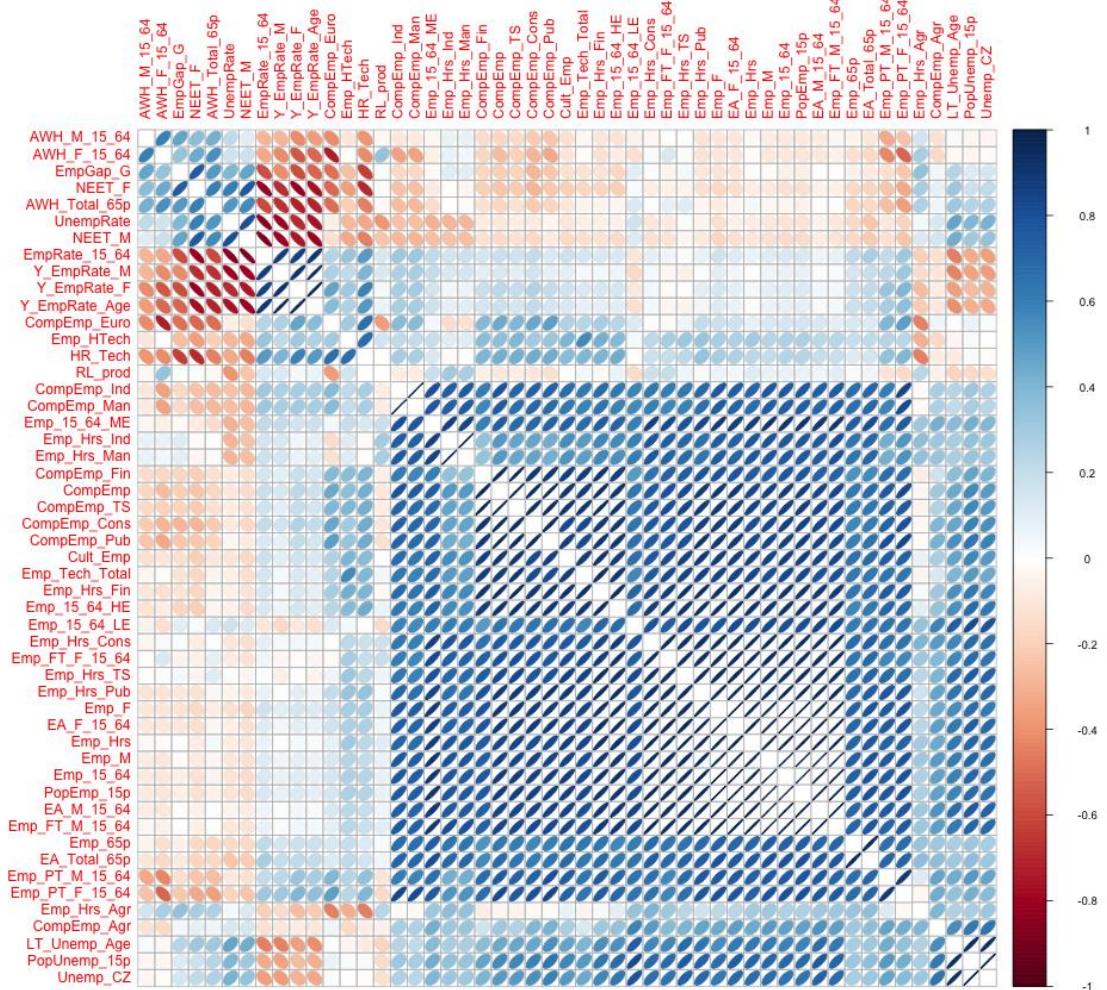


Figura 23 – Correlação entre variáveis

Matriz de correlação:

Devido às diferenças entre as unidades de medida nas variáveis originais, optamos pela matriz de correlação, em vez da matriz de variância, que pode ser observada nas Figuras 24, 25 e 26:

	EmpGap_G	Emp_ITech	Emp_Hrs_Agr	Emp_Hrs_Ind	Emp_Hrs_Man	Emp_Hrs_Cons	Emp_Hrs_TS	Emp_Hrs_Fin	Emp_Pub	EmpRate_15_64	UnempRate	CompEmp_Euro	CompEmp_Agr	CompEmp_Ind	CompEmp_Man	CompEmp_Cons	CompEmp_TS		
EmpGap_G	1.000	-0.253	0.008	0.245	0.195	0.085	0.111	0.039	-0.116	-0.582	0.392	-0.484	-0.205	0.442	-0.165	-0.157	-0.288	-0.184	
Emp_ITech	0.253	1.000	-0.291	0.296	0.296	0.275	0.761	0.961	0.988	0.924	0.967	-0.299	0.329	0.253	0.212	0.295	0.301		
Emp_Hrs_Agr	0.008	-0.291	1.000	-0.296	0.296	0.275	0.761	0.961	0.988	0.924	0.967	-0.299	0.329	0.253	0.212	0.295	0.301		
Emp_Hrs_Ind	0.195	0.103	0.775	0.358	1.000	0.995	0.788	0.692	0.561	0.653	0.160	-0.288	-0.149	0.503	0.318	0.730	0.464	0.462	
Emp_Hrs_Man	0.095	0.103	0.761	0.340	0.995	1.000	0.762	0.675	0.552	0.638	0.177	-0.286	-0.128	0.501	0.320	0.751	0.463	0.456	
Emp_Hrs_Cons	0.011	0.223	0.961	0.404	0.780	0.762	1.000	0.929	0.837	0.911	0.882	-0.110	0.008	0.751	0.497	0.623	0.591	0.770	0.752
Emp_Hrs_TS	0.039	0.318	0.980	0.239	0.692	0.675	0.929	1.000	0.925	0.958	0.813	0.018	0.059	0.813	0.405	0.621	0.586	0.776	0.839
Emp_Hrs_Fin	-0.116	0.425	0.924	0.525	0.561	0.552	0.837	0.925	1.000	0.948	0.120	-0.064	0.286	0.937	0.304	0.677	0.641	0.884	0.956
Emp_Hrs_Pub	-0.097	0.325	0.967	0.163	0.653	0.638	0.911	0.958	0.949	1.000	0.888	-0.041	0.280	0.890	0.463	0.696	0.657	0.867	0.882
EmpRate_15_64	-0.052	-0.275	-0.075	-0.280	-0.280	-0.280	-0.286	-0.118	0.810	-0.041	-0.849	-1.000	-0.299	-0.253	-0.212	-0.212	-0.212	-0.212	-0.212
UnempRate	0.392	-0.275	-0.075	-0.275	-0.280	-0.280	-0.286	-0.118	0.810	-0.041	-0.849	-1.000	-0.299	-0.253	-0.212	-0.212	-0.212	-0.212	-0.212
CompEmp_Euro	-0.484	0.299	0.074	-0.449	-0.149	-0.128	0.003	0.059	0.286	0.293	0.248	-0.063	1.000	0.443	0.058	0.371	0.370	0.459	0.399
CompEmp_Agr	-0.298	0.353	0.840	-0.048	0.503	0.581	0.751	0.813	0.937	0.890	0.186	-0.097	0.443	1.000	0.305	0.781	0.750	0.957	0.986
CompEmp_Ind	0.042	-0.162	0.441	0.393	0.318	0.320	0.497	0.495	0.384	0.463	-0.126	0.139	0.588	0.305	1.000	0.266	0.394	0.254	0.701
CompEmp_Man	-0.166	0.212	0.702	-0.014	0.730	0.751	0.621	0.621	0.677	0.695	0.381	-0.243	0.371	0.781	0.280	1.000	0.998	0.729	0.701
CompEmp_Cons	-0.167	0.208	0.668	-0.021	0.720	0.747	0.591	0.586	0.641	0.657	0.313	-0.252	0.370	0.750	0.266	0.999	1.000	0.698	0.666
CompEmp_TS	-0.184	0.391	0.843	-0.042	0.462	0.456	0.752	0.839	0.882	0.882	0.162	-0.072	0.399	0.986	0.254	0.781	0.666	0.936	1.000
CompEmp_Fin	-0.178	0.346	0.840	-0.042	0.462	0.456	0.752	0.836	0.882	0.882	0.162	-0.072	0.399	0.986	0.254	0.781	0.666	0.936	1.000
CompEmp_Pub	-0.239	0.321	0.822	-0.028	0.432	0.437	0.744	0.795	0.995	0.912	0.128	-0.031	0.381	0.988	0.254	0.781	0.666	0.936	1.000
Cult_Emp	-0.136	0.403	0.862	0.071	0.541	0.538	0.769	0.859	0.917	0.883	0.176	-0.196	0.251	0.897	0.271	0.662	0.630	0.825	0.907
HR_Tech	0.620	0.664	0.224	-0.449	-0.018	-0.889	0.176	0.227	0.401	0.334	0.490	0.321	0.655	0.425	-0.875	0.282	0.272	0.449	0.424
RL_prod	0.019	0.155	0.191	0.231	0.291	0.271	0.175	0.058	0.024	0.138	0.367	-0.093	-0.856	-0.075	-0.078	-0.102	-0.072	-0.072	-0.072
Y_EmpRate_M	-0.413	0.212	0.008	-0.158	0.130	0.141	0.035	-0.063	0.046	0.016	0.079	-0.079	-0.269	0.133	0.261	0.272	0.153	0.096	0.096
Y_EmpRate_F	-0.585	0.302	0.046	-0.269	0.084	0.061	0.043	-0.009	0.131	0.098	0.086	-0.056	-0.465	0.225	-0.074	0.284	0.254	0.192	0.143
Y_EmpRate_Age	-0.528	0.253	0.016	-0.219	0.084	0.096	0.032	-0.044	0.082	0.032	0.077	-0.095	-0.466	0.225	-0.074	0.284	0.204	0.143	0.143
Emp_15_64	-0.063	0.263	0.983	0.263	0.767	0.756	0.954	0.946	0.914	0.969	0.124	-0.111	0.140	0.878	0.477	0.764	0.859	0.862	0.862
Emp_15_65p	-0.058	0.194	0.710	0.263	0.630	0.630	0.838	0.878	0.929	0.964	0.125	-0.115	0.140	0.869	0.466	0.761	0.859	0.862	0.862
Emp_M	-0.116	0.240	0.808	0.263	0.768	0.756	0.945	0.986	0.923	0.969	0.128	-0.100	0.141	0.870	0.469	0.760	0.859	0.862	0.862
Emp_F	-0.119	0.268	0.975	0.228	0.732	0.732	0.943	0.938	0.925	0.979	0.161	-0.133	0.174	0.894	0.447	0.765	0.859	0.862	0.862
PopEmp_15p	-0.064	0.263	0.982	0.268	0.769	0.759	0.953	0.945	0.913	0.968	0.129	-0.116	0.148	0.877	0.476	0.767	0.857	0.861	0.861
PopEmp_15sp	0.120	0.028	0.791	0.272	0.343	0.329	0.684	0.756	0.623	0.727	-0.314	0.309	0.062	0.558	0.615	0.318	0.285	0.588	0.562
Emp_15_64_KE	0.092	0.038	0.885	0.333	0.597	0.604	0.784	0.811	0.707	0.772	-0.184	0.163	0.110	0.635	0.575	0.555	0.633	0.628	0.628
Emp_15_64_ME	-0.027	0.192	0.876	0.319	0.856	0.845	0.784	0.743	0.834	0.820	-0.187	0.165	0.073	0.737	0.420	0.373	0.607	0.690	0.690
Emp_15_64_HE	-0.178	0.380	0.912	0.555	0.537	0.537	0.730	0.873	0.911	0.938	0.124	-0.056	0.230	0.922	0.336	0.585	0.899	0.917	0.917
EA_M_15_64	0.080	0.222	0.981	0.304	0.757	0.747	0.955	0.923	0.892	0.965	0.120	-0.046	0.246	0.924	0.359	0.731	0.832	0.834	0.834
EA_F_15_64	0.084	0.263	0.983	0.304	0.757	0.747	0.955	0.923	0.892	0.965	0.120	-0.046	0.246	0.924	0.359	0.731	0.832	0.834	0.834
EA_F_15_65p	0.084	0.232	0.972	0.304	0.757	0.747	0.955	0.923	0.892	0.965	0.120	-0.046	0.246	0.924	0.359	0.731	0.832	0.834	0.834
EA_F_15_64	-0.092	0.232	0.712	-0.026	0.517	0.528	0.848	0.746	0.769	0.769	-0.165	0.165	0.036	0.882	0.436	0.830	0.813	0.791	0.795
NEET_M	0.462	-0.330	-0.184	0.123	-0.244	-0.255	-0.135	-0.040	-0.115	-0.088	-0.168	-0.088	0.088	0.809	0.418	-0.270	-0.288	-0.164	-0.128
NEET_F	0.749	-0.358	-0.073	0.284	-0.016	-0.831	-0.074	-0.045	-0.178	-0.128	-0.162	-0.082	0.095	0.808	0.418	-0.248	-0.255	-0.298	-0.219
Unemp_CZ	0.146	-0.085	0.662	0.318	0.316	0.308	0.655	0.715	0.569	0.645	-0.358	0.415	0.019	0.493	0.642	0.263	0.230	0.539	0.504
LT_Unemp_Age	0.247	-0.031	0.617	0.262	0.286	0.272	0.599	0.686	0.536	0.638	-0.442	0.478	0.080	0.448	0.549	0.245	0.214	0.453	0.468

Figura 24 – Matriz de correlação

	CompEmp_Fin	CompEmp_Pub	Cult_Emp	HR_Tech	RL_prod	Y_EmpRate_M	Y_EmpRate_Age	Emp_15_64	Emp_15_65p	Emp_M	Emp_F	PopEmp_15p	PopUnemp_15p	Emp_15_64_ME	Emp_15_64_HE	EA_M_15_64	EA_F_15_64	EA_Total_65p		
EmpGap_G	-0.178	-0.230	-0.136	-0.620	-0.199	-0.413	-0.585	-0.528	-0.063	-0.058	-0.016	-0.119	-0.064	-0.120	0.092	-0.027	-0.170	0.009	-0.092	-0.083
Emp_ITech	0.367	0.321	0.403	0.664	0.155	0.212	0.302	0.253	0.263	0.194	0.240	0.288	0.263	0.082	0.038	0.192	0.336	0.225	0.219	
Emp_Hrs_Agr	0.744	0.822	0.862	0.224	0.101	0.006	0.040	0.016	0.983	0.923	0.701	0.701	0.085	0.070	0.080	0.912	0.381	0.977	0.770	
Emp_Hrs_Ind	-0.077	-0.020	0.071	-0.449	-0.231	-0.158	-0.269	-0.219	0.263	0.325	0.301	0.228	0.268	0.272	0.333	0.092	0.304	0.236	0.360	
Emp_Hrs_Man	0.037	0.432	0.541	-0.018	0.291	0.130	0.849	0.884	0.884	0.884	0.767	0.730	0.769	0.759	0.057	0.343	0.555	0.555	0.691	
Emp_Hrs_Cons	0.159	0.427	0.538	-0.018	0.291	0.130	0.849	0.884	0.884	0.884	0.767	0.730	0.769	0.759	0.057	0.343	0.555	0.555	0.691	
Emp_Hrs_TS	0.736	0.795	0.850	-0.227	0.058	-0.463	-0.069	-0.044	0.946	0.646	0.845	0.938	0.945	0.756	0.011	0.874	0.952	0.952	0.782	
Emp_Hrs_Fin	0.981	0.905</																		

	AWH_Tech_Total	AWH_M_15_64	AWH_F_15_64	AWH_Total_65p	Emp_FT_M_15_64	Emp_FT_F_15_64	Emp_PT_M_15_64	Emp_PT_F_15_64	NEET_M	NEET_F	Unemp_CZ	LT_Unemp_Age
EmpGap_G	-0.113	0.466	0.335	0.503	0.001	-0.044	-0.139	-0.223	0.462	0.749	0.146	0.247
Emp_HTech	0.558	-0.189	-0.009	-0.210	0.233	0.282	0.243	0.232	-0.330	-0.358	-0.005	-0.031
Emp_Hrs	0.837	-0.029	-0.016	0.036	0.984	0.959	0.710	0.712	-0.104	-0.073	0.662	0.617
Emp_Hrs_Agr	-0.041	0.151	0.270	0.264	0.310	0.311	0.112	-0.026	0.123	0.284	0.310	0.262
Emp_Hrs_Ind	0.497	0.078	0.072	-0.006	0.803	0.738	0.426	0.517	-0.244	-0.016	0.316	0.286
Emp_Hrs_Man	0.488	0.075	0.049	-0.017	0.791	0.717	0.423	0.528	-0.255	-0.031	0.300	0.272
Emp_Hrs_Cons	0.756	-0.035	0.025	0.013	0.962	0.950	0.658	0.648	-0.135	-0.074	0.655	0.569
Emp_Hrs_TS	0.845	0.002	0.017	0.108	0.948	0.941	0.675	0.657	-0.040	-0.045	0.715	0.686
Emp_Hrs_Fin	0.927	-0.103	-0.122	-0.044	0.887	0.875	0.740	0.746	-0.115	-0.178	0.569	0.536
Emp_Hrs_Pub	0.853	-0.115	-0.116	-0.038	0.955	0.923	0.761	0.769	-0.088	-0.128	0.685	0.638
EmpRate_15_64	0.149	-0.284	-0.324	-0.597	0.079	0.057	0.203	0.306	-0.878	-0.802	-0.350	-0.442
UnempRate	-0.063	0.215	0.166	0.511	-0.092	-0.082	-0.097	-0.188	0.808	0.595	0.415	0.470
CompEmp_Euro	0.236	-0.417	-0.724	-0.494	0.086	-0.018	0.397	0.490	-0.139	-0.508	0.019	0.000
CompEmp	0.874	-0.176	-0.268	-0.178	0.839	0.793	0.758	0.822	-0.149	-0.238	0.493	0.448
CompEmp_Agr	0.137	-0.120	-0.164	0.851	0.488	0.381	0.465	0.436	0.148	0.089	0.642	0.549
CompEmp_Ind	0.615	-0.092	-0.342	-0.268	0.755	0.605	0.605	0.830	-0.270	-0.248	0.263	0.245
CompEmp_Man	0.584	-0.089	-0.348	-0.275	0.723	0.569	0.579	0.813	-0.280	-0.255	0.230	0.214
CompEmp_Cons	0.811	-0.213	-0.287	-0.211	0.820	0.787	0.719	0.791	-0.164	-0.290	0.539	0.453
CompEmp_TS	0.907	-0.151	-0.198	-0.121	0.825	0.809	0.728	0.758	-0.128	-0.219	0.504	0.460
CompEmp_Fin	0.865	-0.150	-0.172	-0.124	0.717	0.728	0.661	0.668	-0.120	-0.203	0.409	0.367
CompEmp_Pub	0.819	-0.233	-0.321	-0.184	0.828	0.769	0.804	0.839	-0.085	-0.223	0.582	0.533
Cult_Emp	0.873	-0.134	-0.107	-0.096	0.835	0.818	0.738	0.722	-0.160	-0.201	0.510	0.464
HR_Tech	0.392	-0.385	-0.413	-0.440	0.185	0.194	0.347	0.397	-0.445	-0.686	-0.013	-0.080
RL_prod	0.047	-0.023	0.332	-0.026	0.095	0.195	-0.116	-0.160	-0.234	0.010	-0.134	-0.185
Y_EmpRate_M	0.033	-0.286	-0.414	-0.669	0.030	-0.032	0.226	0.308	-0.793	-0.679	-0.367	-0.430
Y_EmpRate_F	0.120	-0.411	-0.558	-0.717	0.046	-0.014	0.321	0.396	-0.748	-0.813	-0.293	-0.357
Y_EmpRate_Age	0.081	-0.361	-0.501	-0.707	0.032	-0.030	0.281	0.355	-0.790	-0.769	-0.330	-0.406
Emp_15_64	0.825	-0.084	-0.100	-0.050	0.994	0.946	0.758	0.790	-0.144	-0.123	0.648	0.583
Emp_65p	0.621	-0.045	-0.132	-0.176	0.706	0.631	0.660	0.664	-0.168	-0.164	0.283	0.275
Emp_M	0.802	-0.072	-0.097	-0.034	0.996	0.935	0.759	0.785	-0.122	-0.087	0.661	0.605
Emp_F	0.844	-0.180	-0.107	-0.075	0.982	0.948	0.757	0.795	-0.175	-0.168	0.616	0.544
PopEmp_15p	0.824	-0.085	-0.102	-0.053	0.993	0.944	0.761	0.792	-0.147	-0.125	0.642	0.579
PopUnemp_15p	0.505	-0.037	-0.033	0.248	0.709	0.682	0.504	0.428	0.307	0.172	0.993	0.932
Emp_15_64_LE	0.569	-0.053	-0.139	0.120	0.819	0.735	0.643	0.630	0.116	0.027	0.781	0.733
Emp_15_64_ME	0.665	-0.039	-0.066	-0.143	0.906	0.821	0.658	0.745	-0.252	-0.087	0.400	0.371
Emp_15_64_HE	0.892	-0.119	-0.077	-0.039	0.893	0.920	0.698	0.694	-0.145	-0.208	0.629	0.538
EA_M_15_64	0.793	-0.070	-0.091	-0.003	0.995	0.937	0.751	0.767	-0.074	-0.056	0.719	0.662
EA_F_15_64	0.833	-0.093	-0.100	-0.033	0.985	0.952	0.750	0.777	-0.122	-0.132	0.689	0.615
EA_Total_65p	0.628	-0.186	-0.160	-0.161	0.767	0.688	0.706	0.709	-0.219	-0.185	0.319	0.306
Emp_Tech_Total	1.000	-0.040	-0.016	-0.046	0.798	0.825	0.633	0.635	-0.137	-0.172	0.449	0.405
AWH_M_15_64	-0.040	1.000	0.597	0.424	-0.039	-0.007	-0.321	-0.243	0.106	0.370	-0.032	0.024
AWH_F_15_64	-0.016	0.597	1.000	0.540	-0.050	0.130	-0.424	-0.518	0.174	0.442	-0.013	-0.021
AWH_Total_65p	-0.046	0.424	0.540	1.000	-0.003	0.081	-0.252	-0.356	0.551	0.595	0.269	0.310
Emp_FT_M_15_64	0.798	-0.039	-0.050	-0.003	1.000	0.952	0.703	0.744	-0.114	-0.063	0.671	0.612
Emp_FT_F_15_64	0.825	-0.007	0.130	0.081	0.952	1.000	0.566	0.566	-0.117	-0.059	0.645	0.566
Emp_PT_M_15_64	0.633	-0.321	-0.424	-0.252	0.703	0.566	1.000	0.890	-0.121	-0.236	0.466	0.431
Emp_PT_F_15_64	0.635	-0.243	-0.518	-0.356	0.744	0.566	0.890	1.000	-0.230	-0.327	0.378	0.342
NEET_M	-0.137	0.186	0.174	0.551	-0.114	-0.117	-0.121	-0.230	1.000	0.752	0.338	0.435
NEET_F	-0.172	0.370	0.442	0.595	-0.063	-0.059	-0.236	-0.327	0.752	1.000	0.210	0.314
Unemp_CZ	0.449	-0.032	-0.013	0.269	0.671	0.645	0.466	0.378	0.338	0.210	1.000	0.939
LT_Unemp_Age	0.405	0.024	-0.021	0.310	0.612	0.566	0.431	0.342	0.435	0.314	0.939	1.000

Figura 26 - Continuação da matriz de correlação

Teste de Bartlett e KMO:

O teste de Bartlett verifica se a matriz de correlação é significativamente diferente de uma matriz identidade em que todas as correlações fora da diagonal principal seriam zero.

O valor de qui-quadrado é muito alto (18 582,57), indicando que as correlações entre as variáveis são significativas. O p-value é igual a 0, confirmando que rejeitamos a hipótese nula de que a matriz de correlação é uma matriz identidade.

Esta análise sugere que há correlações suficientes entre as variáveis para justificar o uso do PCA.

```
> cortest.bartlett(correlation)
$chisq
[1] 18582.57

$p.value
[1] 0

$df
[1] 1275
```

Figura 27 – Teste de Bartlett

O KMO mede a adequação da amostra para PCA, avaliando o grau em que as variáveis estão correlacionadas e podem ser agrupadas em componentes principais.

> KMO(correlation)													
Kaiser-Meyer-Olkin factor adequacy													
Call: KMO(r = correlation)													
Overall MSA = 0.82													
MSA for each item =													
EmpGap_G	Emp_HTech	Emp_Hrs	Emp_Hrs_Agr	Emp_Hrs_Ind	Emp_Hrs_Man	Emp_Hrs_Cons	Emp_Hrs_TS	Emp_Hrs_Fin	Emp_Hrs_Pub	EmpRate_15_64			
0.83	0.78	0.80	0.34	0.70	0.92	0.78	0.79	0.79	0.80	0.87			
UnempRate	CompEmp_Euro	CompEmp	CompEmp_Agr	CompEmp_Ind	CompEmp_Man	CompEmp_Cons	CompEmp_TS	CompEmp_Fin	CompEmp_Pub	Cult_Emp			
0.83	0.84	0.78	0.51	0.73	0.93	0.77	0.78	0.74	0.78	0.97			
HR_Tech	RL_prod	Y_EmpRate_M	Y_EmpRate_F	Y_EmpRate_Age	Emp_15_64	Emp_65p	Emp_M	Emp_F	PopEmp_15p	PopUnemp_15p			
0.85	0.51	0.61	0.62	0.61	0.87	0.87	0.92	0.92	0.92	0.75			
Emp_15_64_LE	Emp_15_64_ME	Emp_15_64_HE	EA_M_15_64	EA_F_15_64	EA_Total_65p	Emp_Tech_Total	AWH_M_15_64	AWH_F_15_64	AWH_Total_65p	Emp_FT_M_15_64			
0.88	0.89	0.90	0.86	0.85	0.83	0.95	0.60	0.78	0.90	0.91			
Emp_FT_F_15_64	Emp_PT_M_15_64	Emp_PT_F_15_64	NEET_M	NEET_F	Unemp_CZ	LT_Unemp_Age							
0.91	0.87	0.88	0.81	0.89	0.94	0.95							

Figura 28 – Teste KMO

O KMO geral é 0,82, considerado ser bom e indicando que os dados têm correlações suficientes para serem analisados por PCA. Esta análise sugere que a matriz de correlações é adequada para identificar padrões subjacentes nos dados.

O KMO individual (MSA) varia entre 0,34 e 0,97, com algumas variáveis apresentando valores baixos que indicam baixa adequação para inclusão na análise.

- “Emp_Hrs_Agr” é a variável com o menor MSA (0,34). Esta variável apresenta uma baixa correlação com as restantes e pode ser considerada para exclusão.
- Outras variáveis com MSA moderadamente baixo: “CompEmp_Agr” com 0,51 e “RL_prod” com 0,51.
- Variáveis com alta adequação (MSA > 0,9): “Emp_Hrs_Man” (0,92), “CompEmp_Man” (0,93), “Cult_Emp” (0,97), “Emp_M” (0,92), “Emp_F” (0,92), “Emp_Tech_Total” (0,95), “Unemp_CZ” (0,94), “LT_Unemp_Age” (0,95), “Emp_15_64_HE” (0,90), “AWH_Total_65p” (0,90), “Emp_FT_M_15_64” (0,91), “Emp_FT_F_15_64” (0,91).

Optamos por manter a variável “Emp_Hrs_Agr” da matriz de correlações, tendo em conta que a variável é o indicador que pretendemos integrar na análise posterior de PCA e Clusters, tal como todas as horas trabalhadas em diferentes setores.

Número de PC:

Para a análise do número ideal de componentes principais usámos os dados normalizados (função *scale*). Os resultados podem ser observados nas Figuras 29, 30 e 31.

Principal Components Analysis																												
Call: principal(r = data_scaled, nfactors = 51, rotate = "none", scores = TRUE)																												
Standardized loadings (pattern matrix) based upon correlation matrix																												
	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17	PC18	PC19	PC20	PC21	PC22	PC23	PC24	PC25	PC26		
EmpGap_G	-0.12	0.67	0.24	-0.05	0.25	-0.01	0.20	-0.04	0.54	-0.03	0.00	0.15	-0.03	0.03	-0.09	0.08	-0.05	-0.13	0.01	0.09	-0.01	-0.08	0.06	0.00	0.00	0.00	0.00	
Emp_HTech	0.32	-0.36	-0.12	0.58	-0.06	-0.06	-0.16	0.46	0.30	0.09	0.05	-0.03	-0.22	0.06	0.02	0.00	0.09	0.05	-0.07	-0.04	0.03	0.00	-0.03	-0.04	-0.05	0.02		
Emp_Hrs	0.97	0.15	0.12	0.07	-0.05	0.01	-0.01	0.03	0.02	-0.05	0.00	0.03	0.01	-0.04	0.01	-0.04	0.01	0.02	-0.04	-0.02	-0.01	-0.01	0.04	0.00				
Emp_Hrs_Agr	0.19	0.43	0.51	-0.31	-0.27	-0.22	0.28	0.09	-0.14	0.09	-0.05	0.35	-0.15	-0.01	0.03	-0.10	0.15	0.01	-0.01	-0.05	-0.02	0.00	-0.02	0.01	0.03	0.00		
Emp_Hrs_Ind	0.72	0.08	0.58	-0.10	0.14	0.07	-0.27	0.01	0.03	-0.09	0.00	0.04	0.01	-0.06	-0.02	0.06	-0.01	0.01	0.10	-0.03	0.04	0.02	-0.05	0.04	-0.05	0.00		
Emp_Hrs_Man	0.71	0.06	0.57	-0.12	0.17	0.08	-0.28	0.03	0.02	-0.09	0.01	0.03	0.00	-0.03	-0.03	0.07	0.01	0.01	0.09	-0.02	0.04	0.02	-0.06	0.04	-0.04	0.01		
Emp_Hrs_Cons	0.92	0.16	0.21	0.02	-0.17	-0.01	-0.01	0.01	-0.01	0.01	-0.03	0.07	-0.02	-0.09	-0.04	0.09	-0.03	0.04	0.05	0.03	-0.02	-0.01	0.00	0.03	0.00	0.06		
Emp_Hrs_TS	0.94	0.21	0.03	0.14	-0.10	0.06	0.02	0.06	0.06	-0.10	0.00	0.01	0.05	-0.02	-0.01	0.01	-0.05	0.01	0.01	-0.05	-0.03	-0.04	-0.02	-0.01	0.07	0.02		
Emp_Hrs_Fin	0.94	0.01	-0.15	0.23	0.03	-0.01	0.07	-0.02	0.03	-0.05	0.04	0.00	0.01	0.00	-0.05	0.02	-0.06	0.02	0.03	-0.05	-0.04	-0.03	-0.04	-0.01	0.08	0.00		
Emp_Hrs_Pub	0.97	0.09	-0.06	0.07	-0.07	-0.01	0.01	0.01	0.01	0.00	-0.05	0.02	-0.07	0.02	-0.03	-0.07	-0.02	-0.02	-0.01	-0.01	-0.01	0.04	-0.04	0.01	-0.03	0.03		
EmpRate_15_64	0.17	-0.88	0.26	0.01	-0.14	0.15	0.11	0.02	-0.04	-0.07	0.10	-0.09	0.03	0.11	0.01	-0.02	0.01	0.04	0.03	-0.01	0.04	0.04	0.10	0.05	0.11			
UnempRate	-0.13	0.74	-0.50	-0.05	0.14	0.00	-0.01	-0.11	0.06	0.01	-0.11	0.00	0.13	-0.12	0.06	0.18	-0.14	-0.03	-0.03	0.00	0.04	0.00	-0.05	0.03	0.07			
CompEmp_Euro	0.25	-0.52	-0.65	-0.15	0.19	-0.03	-0.14	0.09	0.02	0.14	-0.08	0.25	0.00	0.07	-0.03	0.02	-0.08	0.13	0.10	0.04	0.14	0.02	0.07	0.05	0.00			
CompEmp	0.92	-0.12	-0.23	0.12	0.16	-0.03	0.07	-0.15	-0.04	0.04	0.01	0.02	-0.03	0.07	0.02	-0.01	0.01	-0.02	0.01	0.01	0.00	-0.01	-0.04	-0.01	-0.01	0.01		
CompEmp_Agr	0.46	0.27	-0.05	-0.54	-0.32	0.03	-0.06	0.00	-0.01	0.43	0.22	-0.11	-0.10	0.08	-0.03	0.17	-0.06	-0.06	0.05	-0.06	-0.03	0.00	-0.01	0.00	0.02	-0.02		
CompEmp_Ind	0.79	-0.22	0.10	-0.17	0.43	0.12	-0.21	-0.04	-0.03	0.02	0.01	0.01	-0.03	0.09	0.07	-0.04	0.11	-0.04	0.02	-0.02	-0.03	0.00	0.00	0.03	-0.03			
CompEmp_Man	0.76	-0.24	0.11	-0.19	0.45	0.13	-0.23	-0.03	-0.03	0.02	0.01	0.01	-0.04	0.11	0.07	-0.04	0.12	-0.05	0.02	-0.02	-0.04	-0.03	0.00	0.01	0.03	-0.03		
CompEmp_Cons	0.90	-0.13	-0.26	0.06	0.03	0.02	0.03	-0.19	-0.11	0.10	-0.02	0.04	-0.05	0.03	0.02	0.00	0.01	0.01	0.05	0.07	0.00	-0.02	0.01	-0.07	0.06			
CompEmp_TS	0.91	-0.08	-0.24	0.22	0.10	-0.04	0.12	-0.14	-0.02	0.00	0.01	0.03	-0.02	0.08	0.00	0.02	0.00	0.00	0.03	-0.01	-0.01	-0.05	-0.01	0.02	0.03			
CompEmp_Fin	0.81	-0.10	-0.28	0.30	0.13	-0.11	0.20	-0.23	-0.04	0.01	0.00	0.05	-0.04	0.10	0.00	0.02	-0.01	0.03	0.06	-0.02	0.04	-0.01	-0.10	0.01	-0.02	0.02		
CompEmp_Pub	0.91	-0.08	-0.32	0.03	0.05	-0.05	0.06	-0.12	-0.03	0.10	0.01	-0.01	0.02	0.00	0.02	-0.07	0.02	0.00	0.04	0.04	0.00	0.03	-0.05	-0.01				
Cult_Emp	0.90	-0.04	-0.10	0.10	0.20	0.04	-0.07	0.13	-0.01	0.00	-0.06	0.03	-0.05	0.03	0.07	-0.07	0.08	-0.10	0.28	-0.10	0.04	0.08	0.12	-0.02	-0.05	0.01		
HR_Tech	0.33	-0.63	-0.38	0.29	-0.16	0.01	-0.23	0.22	0.04	0.10	0.05	0.16	0.06	-0.16	0.08	-0.08	-0.04	-0.10	0.08	0.10	-0.13	0.03	-0.02	0.03	0.03	0.01		
RL_prod	0.02	-0.05	0.57	0.31	-0.33	-0.41	-0.30	-0.17	0.13	0.18	-0.09	0.02	0.28	0.21	0.01	-0.02	0.04	-0.01	-0.01	0.00	0.00	0.01	0.02	-0.02	0.01	0.02		
Y_EmpRate_M	0.12	-0.86	0.26	-0.14	-0.11	0.11	0.19	-0.12	0.20	-0.03	0.00	0.00	-0.02	-0.02	0.13	0.05	0.04	0.05	0.04	0.04	0.01	0.02	-0.03	-0.01	0.01			
Y_EmpRate_F	0.18	-0.91	0.03	-0.14	-0.15	0.14	0.17	-0.06	0.14	-0.04	0.01	0.00	0.03	0.01	0.12	0.11	0.05	0.04	0.03	-0.06	0.00	-0.02	-0.01	0.08	0.00	-0.03		
Y_EmpRate_Age	0.14	-0.91	0.14	-0.15	-0.13	0.10	0.17	-0.06	0.14	-0.04	0.01	0.00	0.04	0.01	0.12	0.08	0.05	0.04	0.04	0.04	0.01	0.02	0.01	0.00	0.01	-0.02		
Emp_15_64	0.99	0.08	-0.09	-0.01	-0.03	0.01	-0.02	-0.03	-0.01	0.00	0.00	0.00	-0.05	-0.01	0.00	0.02	-0.03	0.01	0.00	0.01	0.02	-0.01	0.00	0.00	0.00			
Emp_65p	0.74	-0.06	0.24	-0.15	0.23	-0.24	0.21	0.35	-0.15	-0.01	-0.11	-0.08	0.08	0.09	0.08	0.07	-0.09	-0.05	0.01	0.05	0.02	0.00	0.02	0.03	-0.01	0.03		
Emp_M	0.98	0.11	0.11	-0.05	-0.02	0.01	-0.02	0.01	0.02	0.00	0.00	0.01	0.00	-0.04	-0.01	-0.02	0.00	0.02	0.01	0.01	0.00	0.02	-0.01	0.00	0.00			
Emp_F	0.99	0.03	0.07	0.03	-0.03	0.00	0.00	-0.03	-0.05	0.00	0.00	0.01	-0.05	-0.01	0.01	0.00	0.02	0.04	0.01	-0.01	0.01	0.02	-0.02	-0.01	0.01			
PopEmp_15p	0.99	0.08	-0.09	-0.01	-0.02	0.00	-0.01	-0.02	-0.01	0.00	0.00	0.00	-0.04	-0.01	0.00	0.00	0.02	-0.03	0.01	0.00	0.02	-0.01	0.00	0.00	0.00			
PopUnemp_15p	0.67	0.50	-0.29	-0.15	-0.34	0.19	-0.04	0.02	0.05	-0.05	-0.08	0.03	0.00	0.08	0.04	-0.04	0.05	-0.02	0.01	0.05	0.02	0.03	0.03	0.05	-0.02	0.01		
Emp_15_64_LE	0.80	0.32	-0.05	-0.28	-0.08	0.13	0.00	0.15	-0.02	-0.16	-0.01	0.00	0.04	0.16	-0.20	0.05	0.05	0.13	0.01	0.06	-0.08	0.04	0.00	-0.08	0.01	0.02		
Emp_15_64_ME	0.87	-0.02	0.35	-0.09	0.11	-0.06	-0.07	-0.07	0.02	0.09	0.02	-0.05	-0.07	-0.14	0.08	-0.09	-0.07	-0.01	-0.02	0.04	0.09	-0.03	0.06	-0.03	0.02	0.06		
Emp_15_64_HE	0.93	0.02	-0.12	0.24	-0.13	0.01	0.02	-0.09	-0.02	0.00	-0.01	0.06	0.05	-0.07	0.01	0.05	0.04	-0.02	0.03	-0.04	0.02	-0.01	0.05	-0.03	-0.08			
EA_M_15_64	0.98	0.16	0.07	-0.05	-0.06	0.03	-0.03	0.02	-0.01	0.01	0.00	-0.01	-0.03	0.00	-0.02	0.00	0.02	-0.01	0.01	0.01	0.00	0.02	-0.01	0.00	0.00			
EA_F_15_64	0.99	0.10	0.02	0.01	-0.07	0.03	-0.01	-0.03	-0.03	0.00	0.00	-0.01	0.01	-0.04	0.01	0.01	0.01	0.00	0.02	0.00	0.00	0.02	-0.02	-0.01	0.01			
EA_Total_65p	0.80	-0.08	0.27	-0.17	0.16	-0.22	0.18	0.30	-0.12	-0.04	0.00	-0.06	0.02	0.07	0.09	0.00	-0.09	-0.04	-0.01	0.05	-0.03	0.01	-0.04	-0.04	0.02			
Emp_Tech_Total	0.86	-0.02	-0.11	0.41	0.09	-0.06	0.08	0.05	0.04	-0.05	-0.03	-0.05	-0.01	0.05	-0.01	0.03	0.01	0.01	-0.10	0.04	0.01	0.04	0.11	0.06	-0.06			
AWH_M_15_64	-0.13	0.42	0.33	0.29	0.30	0.51	0.25	0.11	0.02																			

Figura 30 – Continuação PCA

Figura 31 – Resumo PCA

A complexidade média das variáveis foi calculada como 2,5, indicando que muitas variáveis contribuem significativamente para mais de um componente principal.

Variância explicada:

> pc51\$loadings

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14	PC15	PC16	PC17	PC18	PC19	PC20	PC21	PC22
EmpGap_G	-0.123	0.669	0.239		0.248		0.202		0.544		0.153							-0.125				
Emp_HTech	0.320	-0.357	-0.120	0.578			-0.157	0.458	0.300				-0.217									
Emp_Hrs	0.970	0.154	0.116																			
Emp_Hrs_Agr	0.187	0.433	0.505	-0.314	-0.271	-0.222	0.278		-0.140		0.350	-0.151			-0.103	0.153						
Emp_Hrs_Ind	0.719		0.576		0.138			-0.270													0.101	
Emp_Hrs_Man	0.712		0.570	-0.120	0.166			-0.281														
Emp_Hrs_Cons	0.923	0.162	0.208		-0.172																	
Emp_Hrs_TS	0.937	0.214		0.135	-0.102			-0.102														
Emp_Hrs_Fin	0.941		-0.150	0.228																		
Emp_Hrs_Pub	0.973																					
EmpRate_15_64	0.175	-0.876	0.265		-0.142	0.153	0.111			0.100			0.112									
UnempRate	-0.125	0.741	-0.501		0.136		0.111			-0.112	0.131	-0.116	0.181	0.178	-0.138							
CompEmp_Euro	0.253	-0.516	-0.651	-0.152	0.187		-0.145		0.138		0.252				0.128						0.135	
CompEmp	0.923	-0.117	-0.230	0.122	0.157			-0.151														
CompEmp_Agr	0.459	0.273		-0.544	-0.319			0.431	0.225	-0.110	-0.103			0.169								
CompEmp_Ind	0.793	-0.220		-0.175	0.431	0.120	-0.213								0.106							
CompEmp_Man	0.761	-0.236	0.110	-0.191	0.449	0.130	-0.226				0.113			0.123								
CompEmp_Cons	0.898	-0.130	-0.258				-0.194	-0.105														
CompEmp_TS	0.907		-0.244	0.221	0.105		0.119	-0.135														
CompEmp_Fin	0.814	-0.104	-0.280	0.299	0.126	-0.105	0.202	-0.229														
CompEmp_Pub	0.912		-0.323				-0.116															
Cult_Emp	0.897		-0.102	0.202			0.130												0.203	-0.101		
HR_Tech	0.330	-0.634	-0.379	0.291	-0.163		-0.231	0.224	0.103		0.163	-0.164								-0.133		
RL_prod		0.575	0.308	-0.325	-0.409	-0.299	-0.172	0.127	0.176		0.276	0.211										
Y_EmpRate_M	0.116	-0.860	0.262	-0.141	-0.111	0.107	0.190	-0.119	0.195				0.130									
Y_EmpRate_F	0.177	-0.912		-0.135	-0.147		0.126	0.106				0.115	0.106									
Y_EmpRate_Age	0.145	-0.909	0.138	-0.146	-0.135		0.166	0.140				0.123										
Emp_15_64	0.990																					
Emp_65p	0.741		0.236	-0.155	0.230	-0.240	0.211	0.349	-0.151			-0.109										
Emp_M	0.983	0.112	0.112																			
Emp_F	0.991																					
PopEmp_15p	0.990																					
PopUnemp_15p	0.675	0.500	-0.293	-0.148	-0.340	0.185																
Emp_15_64_LE	0.795	0.320		-0.279		0.131		0.155		-0.161			0.158	-0.197		0.134						
Emp_15_64_ME	0.873		0.354		0.113								-0.138									
Emp_15_64_HE	0.929		-0.123	0.240	-0.132																	
EA_M_15_64	0.978	0.161																				
EA_F_15_64	0.987																					
EA_Total_65p	0.796		0.268	-0.174	0.160	-0.222	0.180	0.298	-0.120													
Emp_Tech_Total	0.856		-0.111	0.408																		
AWH_M_15_64	-0.135	0.422	0.326	0.286	0.296	0.513	0.251	0.187		0.350	-0.135	0.161										
AWH_F_15_64	-0.180	0.524	0.450	0.554					-0.173	0.133	-0.176											
AWH_Total_65p	-0.112	0.761		0.225		0.129			-0.126	0.482	0.173	0.141	0.126									
Emp_FT_M_15_64	0.974	0.145	0.139																			
Emp_FT_F_15_64	0.917	0.186	0.161	0.201	-0.152																	
Emp_PT_M_15_64	0.796	-0.149	-0.187	-0.244		-0.204	0.197		0.163	0.143		0.157	-0.122	-0.117		0.111						
Emp_PT_F_15_64	0.831	-0.267	-0.146	-0.276	0.207				0.121		0.123											
NEET_M	-0.170	0.780	-0.389	-0.108	0.159	-0.244							0.164			0.190			-0.119			
NEET_F	-0.194	0.837			0.232	-0.169		-0.156	0.237			-0.105	0.126							0.192		
Unemp_CZ	0.628	0.533	-0.278	-0.174	-0.379	0.167																
LT_Unemp_Age	0.568	0.591	-0.296	-0.178	-0.248	0.146			0.147	-0.113		0.133	-0.145									

Figura 32 – Pesos dos componentes principais

	PC23	PC24	PC25	PC26	PC27	PC28	PC29	PC30	PC31	PC32	PC33	PC34	PC35	PC36	PC37	PC38	PC39	PC40	PC41	PC42	PC43	PC44
EmpGap_G																						
Emp_HTech	1																					
Emp_Hrs		1																				
Emp_Hrs_Agr			1																			
Emp_Hrs_Ind				1																		
Emp_Hrs_Man					1																	
Emp_Hrs_Cons						1																
Emp_Hrs_TS							1															
Emp_Hrs_Fin								1														
Emp_Hrs_Pub									1													
EmpRate_15_64										1												
UnempRate											1											
CompEmp_Euro												1										
CompEmp													1									
CompEmp_Agr														1								
CompEmp_Ind															1							
CompEmp_Man																1						
CompEmp_Cons																	1					
CompEmp_TS																		1				
CompEmp_Fin																			1			
CompEmp_Pub																				1		
Cult_Emp																					1	
HR_Tech																						
RL_prod																						
Y_EmpRate_M																						
Y_EmpRate_F																						
Y_EmpRate_Age																						
Emp_15_64																						
Emp_65p																						
Emp_M																						
Emp_F																						
PopEmp_15p																						
PopUnemp_15p																						
Emp_15_64_LE																						
Emp_15_64_ME																						
Emp_15_64_HE																						
EA_M_15_64																						
EA_F_15_64																						
EA_Total_65p																						
Emp_Tech_Total																						
AWH_M_15_64																						
AWH_F_15_64																						
AWH_Total_65p																						
Emp_FT_M_15_64																						
Emp_FT_F_15_64																						
Emp_PT_M_15_64																						
Emp_PT_F_15_64																						
NEET_M																						
NEET_F																						
Unemp_CZ																						
LT_Unemp_Age																						

Figura 33 – Continuação dos pesos dos componentes principais

Os componentes principais 25, 27, 28 e todos os restantes a partir do trigésimo não explicam uma única variável input e têm a variância explicada igual a 0.

Comunalidades:

Comunalidade é a percentagem de informação de cada variável que é retida nos componentes escolhidos, deste modo as comunalidades apresentaram o valor de 1 para todas as variáveis originais.

> pc51\$community	EmpGap_G	Emp_HTech	Emp_Hrs	Emp_Hrs_Agr	Emp_Hrs_Ind	Emp_Hrs_Man	Emp_Hrs_Cons	Emp_Hrs_TS	Emp_Hrs_Fin	Emp_Hrs_Pub	EmpRate_15_64
	1	1	1	1	1	1	1	1	1	1	1
UnempRate	1	1	1	1	1	1	1	1	1	1	1
CompEmp_Euro											
CompEmp											
CompEmp_Agr											
CompEmp_Ind											
CompEmp_Man											
CompEmp_Cons											
CompEmp_TS											
CompEmp_Fin											
CompEmp_Pub											
Cult_Emp											
HR_Tech											
RL_prod											
Y_EmpRate_M											
Y_EmpRate_F											
Y_EmpRate_Age											
Emp_15_64											
Emp_65p											
Emp_M											
Emp_F											
PopEmp_15p											
PopUnemp_15p											
Emp_15_64_LE											
Emp_15_64_ME											
Emp_15_64_HE											
EA_M_15_64											
EA_F_15_64											
EA_Total_65p											
Emp_Tech_Total											
AWH_M_15_64											
AWH_F_15_64											
AWH_Total_65p											
Emp_FT_M_15_64											
Emp_PT_M_15_64											
Emp_PT_F_15_64											
NEET_M											
NEET_F											
Unemp_CZ											
LT_Unemp_Age											

Figura 34 – Comunalidades 51 componentes principais

Extração dos componentes principais:

PC4 vs PC5

Principal Components Analysis												
Call: principal(r = data_scaled, nfactors = 4, rotate = "none")												
Standardized loadings (pattern matrix) based upon correlation matrix												
	PC1	PC2	PC3	PC4	h2	u2	com	PC1	PC2	PC3	PC4	
EmpGap_G	-0.12	0.67	0.24	-0.05	0.52	0.4785	1.3	EmpGap_G	-0.12	0.67	0.24	-0.05
Emp_HTech	0.32	-0.36	-0.12	0.58	0.58	0.4213	2.4	Emp_HTech	0.32	-0.36	-0.12	0.58
Emp_Hrs	0.97	0.15	0.12	0.07	0.98	0.0173	1.1	Emp_Hrs	0.97	0.15	0.12	0.07
Emp_Hrs_Agr	0.19	0.43	0.51	-0.31	0.58	0.4232	3.0	Emp_Hrs_Agr	0.19	0.43	0.51	-0.31
Emp_Hrs_Ind	0.72	0.08	0.58	-0.10	0.86	0.1361	2.0	Emp_Hrs_Ind	0.72	0.08	0.58	-0.10
Emp_Hrs_Man	0.71	0.06	0.57	-0.12	0.85	0.1503	2.0	Emp_Hrs_Man	0.71	0.06	0.57	-0.12
Emp_Hrs_Cons	0.92	0.16	0.21	0.02	0.92	0.0783	1.2	Emp_Hrs_Cons	0.92	0.16	0.21	0.02
Emp_Hrs_TS	0.94	0.21	0.03	0.14	0.94	0.0571	1.2	Emp_Hrs_TS	0.94	0.21	0.03	0.14
Emp_Hrs_Fin	0.94	0.01	-0.15	0.23	0.96	0.0406	1.2	Emp_Hrs_Fin	0.94	0.01	-0.15	0.23
Emp_Hrs_Pub	0.97	0.09	-0.06	0.07	0.96	0.0370	1.0	Emp_Hrs_Pub	0.97	0.09	-0.06	0.07
EmpRate_15_64	0.17	-0.88	0.26	0.01	0.87	0.1322	1.3	EmpRate_15_64	0.17	-0.88	0.26	0.01
UnempRate	-0.13	0.74	-0.50	-0.05	0.82	0.1822	1.8	UnempRate	-0.13	0.74	-0.50	-0.05
CompEmp_Euro	0.25	-0.52	-0.65	-0.15	0.78	0.2226	2.4	CompEmp_Euro	0.25	-0.52	-0.65	-0.15
CompEmp	0.92	-0.12	-0.23	0.12	0.93	0.0663	1.2	CompEmp	0.92	-0.12	-0.23	0.12
CompEmp_Agr	0.46	0.27	-0.05	-0.54	0.58	0.4167	2.5	CompEmp_Agr	0.46	0.27	-0.05	-0.54
CompEmp_Ind	0.79	-0.22	0.10	-0.17	0.72	0.2835	1.3	CompEmp_Ind	0.79	-0.22	0.10	-0.17
CompEmp_Man	0.76	-0.24	0.11	-0.19	0.68	0.3162	1.4	CompEmp_Man	0.76	-0.24	0.11	-0.19
CompEmp_Cons	0.90	-0.13	-0.26	0.06	0.89	0.1070	1.2	CompEmp_Cons	0.90	-0.13	-0.26	0.06
CompEmp_TS	0.91	-0.08	-0.24	0.22	0.94	0.0633	1.3	CompEmp_TS	0.91	-0.08	-0.24	0.22
CompEmp_Fin	0.81	-0.10	-0.28	0.30	0.84	0.1596	1.6	CompEmp_Fin	0.81	-0.10	-0.28	0.30
CompEmp_Pub	0.91	-0.08	-0.32	0.03	0.94	0.0577	1.3	CompEmp_Pub	0.91	-0.08	-0.32	0.03
Cult_Emp	0.90	-0.04	-0.10	0.20	0.86	0.1423	1.1	Cult_Emp	0.90	-0.04	-0.10	0.20
HR_Tech	0.33	-0.63	-0.38	0.29	0.74	0.2605	2.7	HR_Tech	0.33	-0.63	-0.38	0.29
RL_prod	0.02	-0.05	0.57	0.31	0.43	0.5719	1.5	RL_prod	0.02	-0.05	0.57	0.31
Y_EmpRate_M	0.12	-0.86	0.26	-0.14	0.84	0.1582	1.3	Y_EmpRate_M	0.12	-0.86	0.26	-0.14
Y_EmpRate_F	0.18	-0.91	0.03	-0.14	0.88	0.1179	1.1	Y_EmpRate_F	0.18	-0.91	0.03	-0.14
Y_EmpRate_Age	0.14	-0.91	0.14	-0.15	0.89	0.1129	1.2	Y_EmpRate_Age	0.14	-0.91	0.14	-0.15
Emp_15_64	0.99	0.08	0.09	-0.01	0.99	0.0071	1.0	Emp_15_64	0.99	0.08	0.09	-0.01
Emp_65p	0.74	-0.06	0.24	-0.15	0.63	0.3686	1.3	Emp_65p	0.74	-0.06	0.24	-0.15
Emp_M	0.98	0.11	0.11	-0.05	0.99	0.0056	1.1	Emp_M	0.98	0.11	0.11	-0.05
Emp_F	0.99	0.03	0.07	0.03	0.99	0.0122	1.0	Emp_F	0.99	0.03	0.07	0.03
PopEmp_15p	0.99	0.08	0.09	-0.01	0.99	0.0058	1.0	PopEmp_15p	0.99	0.08	0.09	-0.01
PopUnemp_15p	0.67	0.50	-0.29	-0.15	0.81	0.1870	2.4	PopUnemp_15p	0.67	0.50	-0.29	-0.15
Emp_15_64_LE	0.80	0.32	-0.05	-0.28	0.82	0.1845	1.6	Emp_15_64_LE	0.80	0.32	-0.05	-0.28
Emp_15_64_ME	0.87	-0.02	0.35	-0.05	0.90	0.1044	1.3	Emp_15_64_ME	0.87	-0.02	0.35	-0.05
Emp_15_64_HE	0.93	0.02	-0.12	0.24	0.94	0.0628	1.2	Emp_15_64_HE	0.93	0.02	-0.12	0.24
EA_M_15_64	0.98	0.16	0.07	-0.05	0.99	0.0102	1.1	EA_M_15_64	0.98	0.16	0.07	-0.05
EA_F_15_64	0.99	0.10	0.02	0.01	0.98	0.0157	1.0	EA_F_15_64	0.99	0.10	0.02	0.01
EA_Total_65p	0.80	-0.08	0.27	-0.17	0.74	0.2565	1.4	EA_Total_65p	0.80	-0.08	0.27	-0.17
Emp_Tech_Total	0.86	-0.02	-0.11	0.41	0.91	0.0878	1.5	Emp_Tech_Total	0.86	-0.02	-0.11	0.41
AWH_M_15_64	-0.13	0.42	0.33	0.29	0.38	0.6164	3.0	AWH_M_15_64	-0.13	0.42	0.33	0.29
AWH_F_15_64	-0.18	0.52	0.45	0.55	0.82	0.1842	3.2	AWH_F_15_64	-0.18	0.52	0.45	0.55
AWH_Total_65p	-0.11	0.76	0.05	0.22	0.65	0.3544	1.2	AWH_Total_65p	-0.11	0.76	0.05	0.22
Emp_FT_M_15_64	0.97	0.15	0.14	-0.02	0.99	0.0116	1.1	Emp_FT_M_15_64	0.97	0.15	0.14	-0.02
Emp_FT_F_15_64	0.92	0.19	0.16	0.20	0.94	0.0581	1.2	Emp_FT_F_15_64	0.92	0.19	0.16	0.20
Emp_PT_M_15_64	0.80	-0.15	-0.19	-0.24	0.75	0.2490	1.4	Emp_PT_M_15_64	0.80	-0.15	-0.19	-0.24
Emp_PT_F_15_64	0.83	-0.27	-0.15	-0.28	0.86	0.1411	1.5	Emp_PT_F_15_64	0.83	-0.27	-0.15	-0.28
NEET_M	-0.17	0.78	-0.39	-0.11	0.80	0.2000	1.6	NEET_M	-0.17	0.78	-0.39	-0.11
NEET_F	-0.19	0.84	0.07	-0.02	0.74	0.2567	1.1	NEET_F	-0.19	0.84	0.07	-0.02
Unemp_CZ	0.63	0.53	-0.28	-0.17	0.79	0.2136	2.5	Unemp_CZ	0.63	0.53	-0.28	-0.17
LT_Unemp_Age	0.57	0.59	-0.30	-0.18	0.79	0.2088	2.7	LT_Unemp_Age	0.57	0.59	-0.30	-0.18

Principal Components Analysis												
Call: principal(r = data_scaled, nfactors = 5, rotate = "none")												
Standardized loadings (pattern matrix) based upon correlation matrix												
	PC1	PC2	PC3	PC4	h2	u2	com	PC1	PC2	PC3	PC4	
SS loadings	26.59	9.04	3.91	2.44				SS loadings	26.59	9.04	3.91	2.44
Proportion Var	0.52	0.18	0.08	0.05				Proportion Var	0.52	0.18	0.08	0.05
Cumulative Var	0.52	0.70	0.78	0.82				Cumulative Var	0.52	0.70	0.78	0.82
Proportion Explained	0.63	0.22	0.09	0.06				Proportion Explained	0.61	0.21	0.09	0.06
Cumulative Proportion	0.63	0.85	0.94	1.00				Cumulative Proportion	0.61	0.81	0.90	0.96

Mean item complexity = 1.6
Test of the hypothesis that 4 components are sufficient.

The root mean square of the residuals (RMSR) is 0.04
with the empirical chi square 1071.54 with prob < 0.54

Fit based upon off diagonal values = 0.99
Fit based upon off diagonal values = 1

Figura 35 – PCA para 4 e 5 componentes principais

Métricas	PC4	PC5
Root mean square of the residuals (RMSR)	0,04	0,03
Complexidade	1,6	1,7
Variância acumulada	82%	86%
Variáveis Chave	“Emp_HTech” (0,58), “CompEmp_Agr” (-0,54), “HR_Tech” (0,29)	“CompEmp_Ind” (0,43), “CompEmp_Man” (0,45), “Emp_Hrs_Agr” (-0,27)

O PC4 e o PC5 contribuem para a compreensão das diferenças setoriais nos padrões de emprego, mas enfatizam aspectos diferentes:

- PC4 destaca a divisão entre os setores de alta tecnologia e agricultura, focando nos avanços tecnológicos;
- PC5 enfatiza as dinâmicas industriais e de fabrico, capturando variações específicas de emprego nesses setores.

Comparando as métricas expostas na tabela, com os cinco componentes principais conseguimos diminuir o RMSR e aumentar a variância acumulada.

Rotação:

Para a rotação usámos o método *varimax*. Considerámos este método o mais adequado, pois maximiza a variância das cargas dentro de cada componente entre as variáveis.

RC1:

- Variância Explicada: 50,7% da variância total.
- Principais Variáveis:
 - Altas cargas positivas em variáveis relacionadas ao emprego geral, como “Emp_Hrs” (0,976), “Emp_15_64” (0,985), “Emp_M” (0,983), “Emp_F” (0,982), “PopEmp_15p” (0,986).
 - Também inclui setores específicos como “CompEmp” (0,899) e “CompEmp_TS” (0,885).

Este componente reflete uma dimensão geral de emprego e participação no mercado de trabalho, capturando tanto o emprego total quanto o emprego por setores.

RC2:

- Variância Explicada: 16,6% da variância total.
- Principais Variáveis:
 - Fortes associações negativas com indicadores relacionados ao desemprego e jovens fora do mercado de trabalho: “UnempRate” (-0,833), “NEET_M” (-0,850), “NEET_F” (-0,841).
 - Altas cargas positivas em “EmpRate_15_64” (0,923) e taxas de

Este componente reflete um contraste entre o emprego ativo e o desemprego/NEETs, destacando a situação de jovens no mercado de trabalho.

RC3:

- Variância Explicada: 6,8% da variância total.
- Principais Variáveis:
- Altas cargas em variáveis relacionadas à composição setorial do emprego:
 - “CompEmp_Euro” (0,749), “CompEmp_Ind” (0,301), e “CompEmp_Man” (0,308).
 - Carga negativa em produtividade relativa (“RL_prod” -0,698).

Este componente captura as diferenças na composição setorial do emprego e sua relação com a produtividade.

RC4:

- Variância Explicada: 6,2% da variância total.
- Principais Variáveis:
 - Altas cargas positivas em setores tecnológicos: “Emp_HTech” (0,621), “HR_Tech” (0,595).
 - Cargas negativas em agricultura: “CompEmp_Agr” (-0,417) e “Emp_Hrs_Agr” (-0,581).

Este componente reflete o contraste entre empregos em setores tecnológicos versus agrícolas.

RC5:

- Variância Explicada: 5,5% da variância total.
- Principais Variáveis:
 - Cargas positivas em indicadores relacionados ao desemprego de longo prazo e população desempregada: “PopUnemp_15p” (0,629), “Unemp_CZ” (0,665), e “LT_Unemp_Age” (0,566).
 - Carga positiva também em agricultura (CompEmp_Agr, 0,569).

Este componente captura aspectos do desemprego estrutural e sua relação com setores específicos como a agricultura.

	RC1	RC2	RC3	RC4	RC5	
EmpGap_G		-0.644	-0.185	-0.329	-0.162	
EmpHTech	0.295	0.285	-0.130	0.621	-0.107	
Emp_Hrs	0.976		-0.109		0.143	
Emp_Hrs_Agr	0.223	-0.176	-0.407	-0.581	0.258	
Emp_Hrs_Ind	0.779		-0.277	-0.395	-0.183	
Emp_Hrs_Man	0.772	0.107	-0.247	-0.409	-0.203	
Emp_Hrs_Cons	0.923		-0.205		0.222	
Emp_Hrs_TS	0.939		-0.113		0.203	
Emp_Hrs_Fin	0.927			0.297		
Emp_Hrs_Pub	0.954			0.113	0.211	
EmpRate_15_64	0.103	0.923			-0.132	
UnempRate		-0.833	0.290		0.214	
CompEmp_Euro	0.152	0.347	0.749	0.325		
CompEmp	0.899	0.102	0.260	0.267		
CompEmp_Agr	0.405		0.137	-0.417	0.569	
CompEmp_Ind	0.808	0.220	0.301	-0.162	-0.289	
CompEmp_Man	0.778	0.232	0.308	-0.183	-0.312	
CompEmp_Cons	0.853	0.145	0.257	0.243	0.143	
CompEmp_TS	0.885		0.191	0.351		
CompEmp_Fin	0.794		0.186	0.433		
CompEmp_Pub	0.866		0.319	0.243	0.167	
Cult_Emp	0.884			0.257		
HR_Tech	0.228	0.552	0.212	0.595	0.100	
RL_prod		0.206	-0.698			
Y_EmpRate_M		0.913			-0.117	
Y_EmpRate_F		0.928	0.179			
Y_EmpRate_Age		0.942	0.105			
Emp_15_64	0.985			0.142		
Emp_65p	0.763	0.140		-0.236	-0.148	
Emp_M	0.983			0.144		
Emp_F	0.982			0.126		
PopEmp_15p	0.986			0.136		
PopUnemp_15p	0.633	-0.354		0.629		
Emp_15_64_LE	0.781	-0.174	0.127	-0.222	0.340	
Emp_15_64_ME	0.899	0.161	-0.102	-0.238		
Emp_15_64_HE	0.902			0.306	0.211	
EA_M_15_64	0.973			0.205		
EA_F_15_64	0.974			0.198		
EA_Total_65p	0.810	0.195		-0.256		
Emp_Tech_Total	0.861			0.419		
AWH_M_15_64		-0.450	-0.340		-0.385	
AWH_F_15_64		-0.462	-0.754	0.100	-0.166	
AWH_Total_65p		-0.740	-0.307			
Emp_FT_M_15_64	0.979			0.140		
Emp_FT_F_15_64	0.927		-0.259		0.170	
Emp_PT_M_15_64	0.739	0.210	0.341		0.208	
Emp_PT_F_15_64	0.787	0.287	0.440			
NEET_M	-0.123	-0.850	0.235		0.173	
NEET_F		-0.841	-0.125	-0.248		
Unemp_CZ	0.586	-0.374			0.665	
LT_Unemp_Age	0.545	-0.470	0.111		0.566	
	RC1	RC2	RC3	RC4	RC5	
SS loadings	25.840	8.450	3.454	3.166	2.811	
Proportion Var	0.507	0.166	0.068	0.062	0.055	
Cumulative Var	0.507	0.672	0.740	0.802	0.857	

Figura 36 – Pesos com 5 componentes principais

Comunalidades PC5:

EmpGap_G	EmpHTech	Emp_Hrs	Emp_Hrs_Agr	Emp_Hrs_Ind	Emp_Hrs_Man	Emp_Hrs_Cons	Emp_Hrs_TS	Emp_Hrs_Fin	Emp_Hrs_Pub		
0.58	0.58	0.99	0.65	0.88	0.88	0.95	0.95	0.96	0.97		
EmpRate_15_64	UnempRate	CompEmp_Euro	CompEmp	CompEmp_Agr	CompEmp_Ind	CompEmp_Man	CompEmp_Cons	CompEmp_TS	CompEmp_Fin		
0.89	0.84	0.81	0.96	0.69	0.90	0.89	0.89	0.95	0.86		
CompEmp_Pub	Cult_Emp	HR_Tech	RL_prod	Y_EmpRate_M	Y_EmpRate_Age	Emp_15_64_TS	EA_F_15_64	EA_Total_65p	Emp_Tech_Total		
0.94	0.86	0.77	0.53	0.85	0.90	0.91	0.99	0.68	0.99		
Emp_F	PopEmp_15p	PopUnemp_15p	Emp_15_64_LE	Emp_15_64_ME	Emp_15_64_HE	EA_M_15_64	EA_F_15_64	EA_Total_65p	Emp_Tech_Total		
0.99	0.99	0.93	0.82	0.91	0.95	0.99	0.99	0.77	0.92		
AWH_M_15_64	AWH_F_15_64	AWH_Total_65p	Emp_FT_M_15_64	Emp_FT_F_15_64	Emp_PT_M_15_64	Emp_PT_F_15_64	NEET_M	NEET_F	Unemp_CZ		
0.47	0.82	0.65	0.99	0.96	0.75	0.90	0.83	0.80	0.93		
LT_Unemp_Age											
0.85											

Figura 37 – Comunalidades com 5 componentes principais

Comunalidades Altas ($\geq 0,90$):

“Emp_Hrs”, “Emp_15_64”, “Emp_M”, “Emp_F”, “PopEmp_15p”, “EA_M_15_64”, “EA_F_15_64”, “Emp_FT_M_15_64”, “CompEmp”, “CompEmp_TS”, “CompEmp_Pub”: Estas variáveis têm comunalidades muito altas, indicando que quase toda a sua variância é explicada pelos componentes principais. Isso sugere que são bem representados no espaço dos componentes, refletindo a sua importância central no modelo.

Comunalidades Moderadas (0,70 – 0,89):

“Emp_HTech”, “CompEmp_Euro”, “CompEmp_Ind”, “CompEmp_Man”, “Cult_Emp”, “HR_Tec”h: Estas variáveis têm uma boa parte da sua variância explicada, mas não tanto quanto as anteriores. Elas ainda são bem representadas, mas podem ter variações que não são totalmente capturadas pelos componentes.

“UnempRate”, “Y_EmpRate_M”, “Y_EmpRate_F”, “Y_EmpRate_Age”: As taxas de emprego e desemprego jovem também estão bem representadas, refletindo o seu impacto significativo nos componentes.

Comunalidades Baixas ($\leq 0,69$):

“RL_prod”, “AWH_M_15_64”, “AWH_Total_65p”: Estas variáveis têm comunalidades mais baixas, indicando que uma parte significativa da sua variância não é explicada pelos componentes principais. Isso sugere que pode haver características únicas ou ruído que não são capturados pelo modelo.

Scores:

```
> round(pc5sc$scores,3)
      PC1     PC2     PC3     PC4     PC5
[1,] -0.265  0.112 -1.981  0.981  0.520
[2,]  0.125 -0.711 -0.825  0.450  0.506
[3,] -0.542 -0.834 -0.650 -0.160 -0.023
[4,] -0.139 -0.822 -0.641  0.312  0.158
[5,] -0.354 -0.814 -1.054  0.978 -0.077
[6,] -0.332 -0.730 -0.215  0.027  0.180
[7,] -0.761 -0.574 -0.897  1.495 -0.346
[8,] -0.488  0.212 -1.240  0.171  0.435
[9,] -0.542  0.042 -1.084  0.329  0.453
[10,] -0.855 -0.332 -0.661  0.060  0.303
[11,] -0.818 -0.189 -1.142  0.578  0.317
[12,] -0.889  1.329  0.381  0.566  0.074
[13,] -0.777  0.407  0.761  0.480 -0.483
[14,] -0.754  0.777  0.971  0.447 -0.200
[15,] -0.723  0.864  0.977  0.383  0.043
[16,]  0.153 -0.090  1.185  1.994 -1.179
[17,] -0.505  0.775  1.473  0.523 -0.262
[18,]  0.036 -0.711  0.325  2.256 -0.325
[19,] -0.311 -0.111  1.521  0.916 -0.265
[20,] -0.374 -0.078  1.392  0.266 -0.098
[21,] -0.533  0.278  1.103  0.276  0.405
[22,] -0.214 -0.018  1.689  0.478  0.076
[23,] -0.052 -0.074  1.425  0.477 -0.131
[24,] -0.413  0.083  1.322  0.386  0.134
[25,] -0.429 -0.065  0.772  0.286  0.042
[26,]  0.485 -1.408 -1.318  1.053 -0.482
[27,] -0.516 -0.764 -0.624 -0.289 -0.494
[28,] -0.194 -0.815 -0.673 -1.086 -0.289
[29,] -0.057 -1.023 -0.601 -0.629 -0.596
[30,] -0.633 -0.761 -0.695 -0.639 -0.456
[31,]  2.268 -1.313  1.201 -1.309  4.551
[32,]  1.049 -1.196  0.348 -0.295  1.825
[33,]  0.673 -1.259  0.623 -0.722  1.476
[34,]  0.437 -1.358  0.550 -0.653  1.355
[35,]  2.724 -1.375  0.598  0.277  3.405
[36,] -0.127 -1.304  0.352 -1.008  0.402
[37,] -0.196 -1.265  0.127 -0.476  0.389
[38,] -0.258 -1.311  0.129 -0.658  0.064
[39,]  0.251 -1.104  0.115 -0.101  0.841
[40,] -0.056 -1.177  0.044 -0.797  0.264
[41,]  0.329 -1.241  0.512 -0.859  0.926
[42,]  1.584 -0.850 -0.965  1.489  0.046
[43,]  0.435 -0.661  0.155 -0.514 -0.495
[44,] -0.552 -0.809 -0.716 -0.349  0.275
[45,]  0.440 -1.120 -0.740  0.683  0.444
[46,]  1.854 -1.005 -0.285  0.389  1.841
[47,] -0.333 -1.089 -0.164 -0.348  0.169
[48,] -0.201 -1.133 -0.074 -0.809  0.285
[49,] -0.103 -0.757  0.015 -0.644 -0.658
[50,]  0.078 -0.968  0.265 -0.373  0.589
[51,]  0.375 -0.913 -0.143 -0.611  0.567
[52,] -0.015 -0.886  0.070 -1.007 -0.057
[53,]  0.634 -0.816  0.503 -1.694  0.657
[54,]  2.294 -0.572  0.240 -0.867  2.525
[55,]  1.910 -0.785 -0.177 -0.111  1.813
```

Figura 38 – Scores das 5 componentes principais

```

> # Cálculo da média
> mean(pc5sc$scores[,1])
[1] 5.229976e-17
> # Cálculo do desvio padrão
> sd(pc5sc$scores[,1])
[1] 1

```

Figura 39 – Média e desvio padrão dos scores

Scores normalizados:

```

> head(normalized_scores)
      PC1       PC2       PC3       PC4       PC5
[1,] -0.05132692  0.03726797 -1.0016867  0.6276366  0.39452953
[2,]  0.02427145 -0.23634018 -0.4171901  0.2877663  0.38413215
[3,] -0.10520683 -0.27748306 -0.3286378 -0.1020192 -0.01747105
[4,] -0.02695314 -0.27343143 -0.3240918  0.1996856  0.12011670
[5,] -0.06858086 -0.27071718 -0.5327939  0.6257142 -0.05876901
[6,] -0.06441240 -0.24294293 -0.1085444  0.0175145  0.13633060

```

Figura 40 – Scores normalizados

Médias dos Scores:

- PC1: -0,0487
- PC2: -0,2106
- PC3: -0,4522
- PC4: 0,2760
- PC5: 0,1598

As médias indicam que, em geral, as observações têm scores ligeiramente negativos em PC1, PC2 e PC3, enquanto apresentam scores positivos em PC4 e PC5.

Desvios Padrão:

- PC1: 0,0401
- PC2: 0,1119
- PC3: 0,2767
- PC4: 0,2774
- PC5: 0,1764

Os desvios padrões mostram a variabilidade dos scores para cada componente. Os componentes PC3 e PC4 têm maior variabilidade, indicando que as observações diferem mais significativamente ao longo dessas dimensões.

PC1 vs PC2

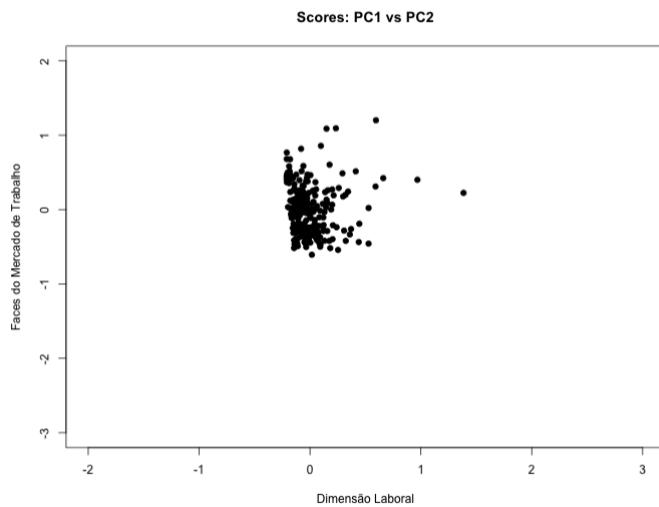


Figura 41 – Gráfico de dispersão PC1 vs PC2

- A concentração de pontos sugere que as observações possuem características semelhantes em relação aos dois componentes principais.
- A dispersão ao longo do eixo X indica que a “Dimensão Laboral” contribui mais para a variação nos dados do que a “Faces do Mercado de Trabalho”.

Anexo C

Clustering

Hierarchical Clustering:

O método utilizado para a separação dos *clusters* foi o Ward.D2, que funciona minimizando a variância dentro dos grupos formados. Esse método procura unir os *clusters* que provocam o menor aumento na soma dos quadrados das distâncias entre observações e o centroide do *cluster*.

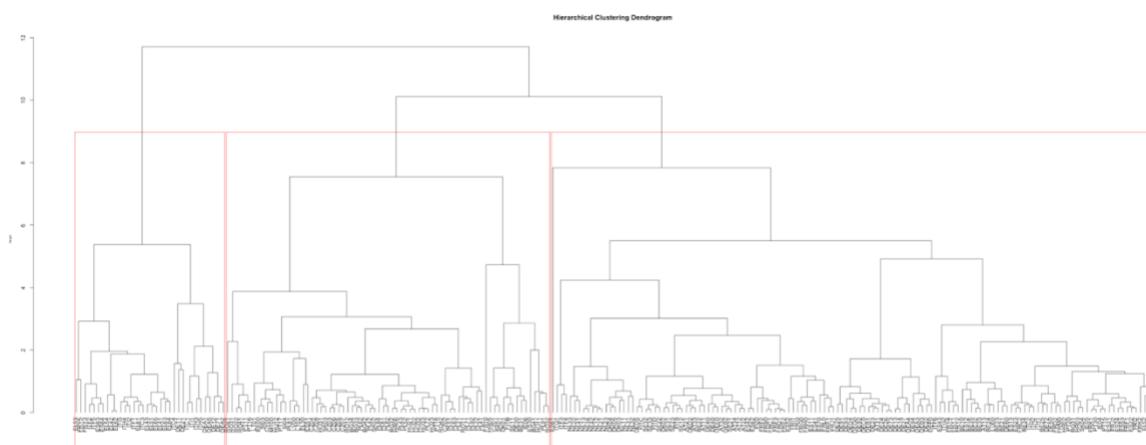


Figura 42 - Hierarchical Clustering

Pode-se observar pelo *Silhouette score* que algumas observações se encontram próximas às bordas dos agrupamentos, o que sugere que, embora os *clusters* estejam bem definidos no geral, ainda há observações com características intermediárias entre dois *clusters*.

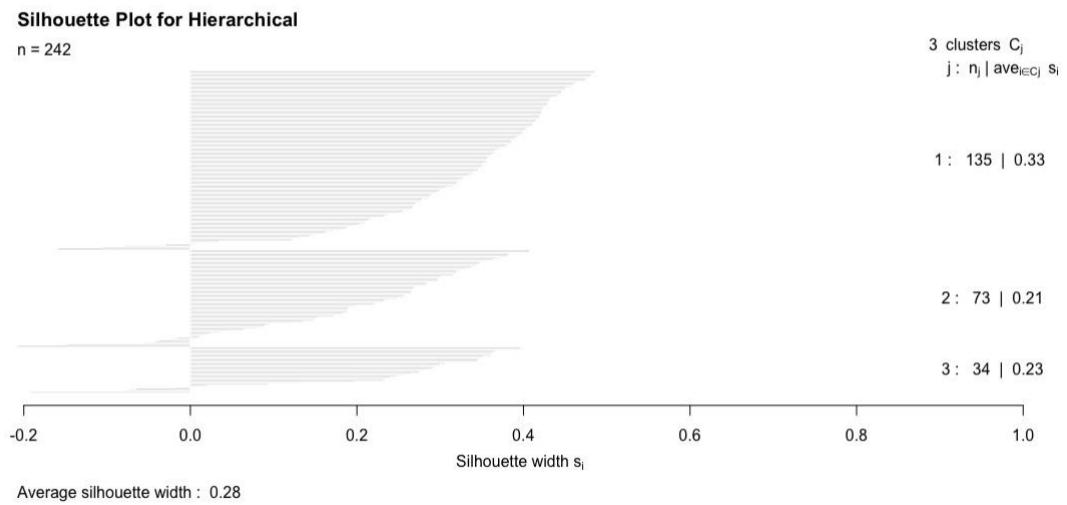


Figura 43 - Silhouette Plot Hierarchical Clustering

Elbow plot para definir número de clusters:

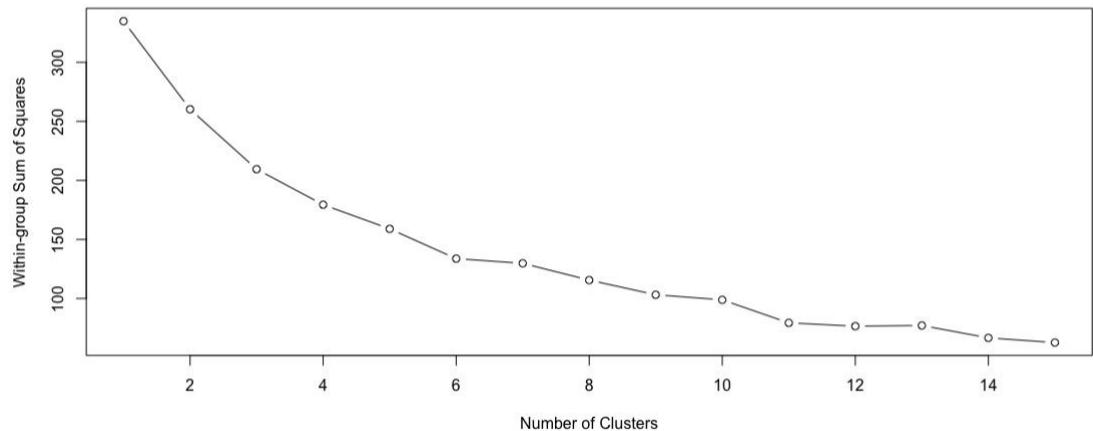


Figura 44 – Elbow Plot

Método utilizado para testar diferentes números de *clusters*, para esse relatório foram testados números de 1 a 15, com o objetivo de determinar o número ideal de *clusters*. Baseado no gráfico a concluir que o número ideal é 4.

K-Means:

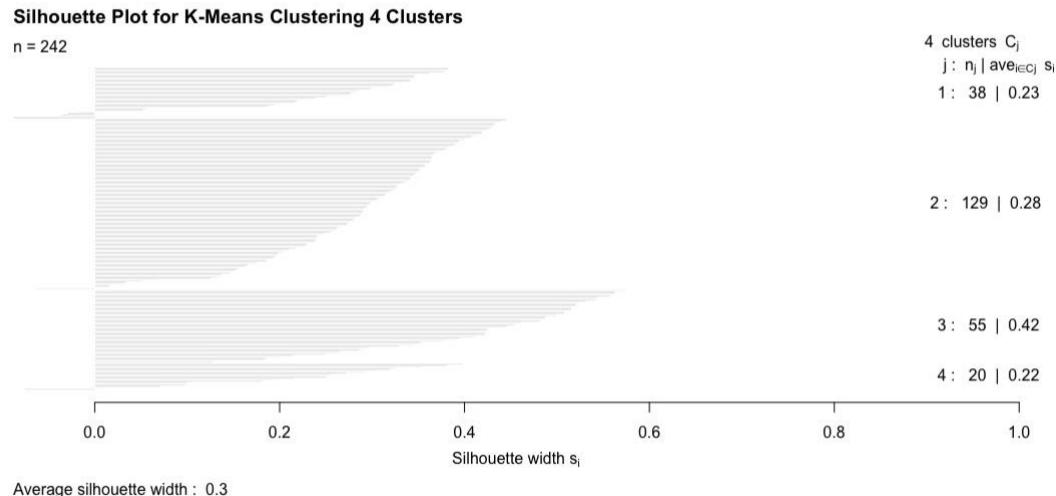


Figura 45 - Silhouette Plot K-Means Clustering

Apesar de ser um método não muito robusto e sensível a outliers, o *K-Means*, para 4 clusters, foi o que obteve o melhor resultado entre os modelos testado para esse relatório. Foram realizadas 25 reiniciações do algoritmo para buscar o centroide ótimo, aquele que vai ter a menor soma dos quadrados *intra-cluster*.

PAM:

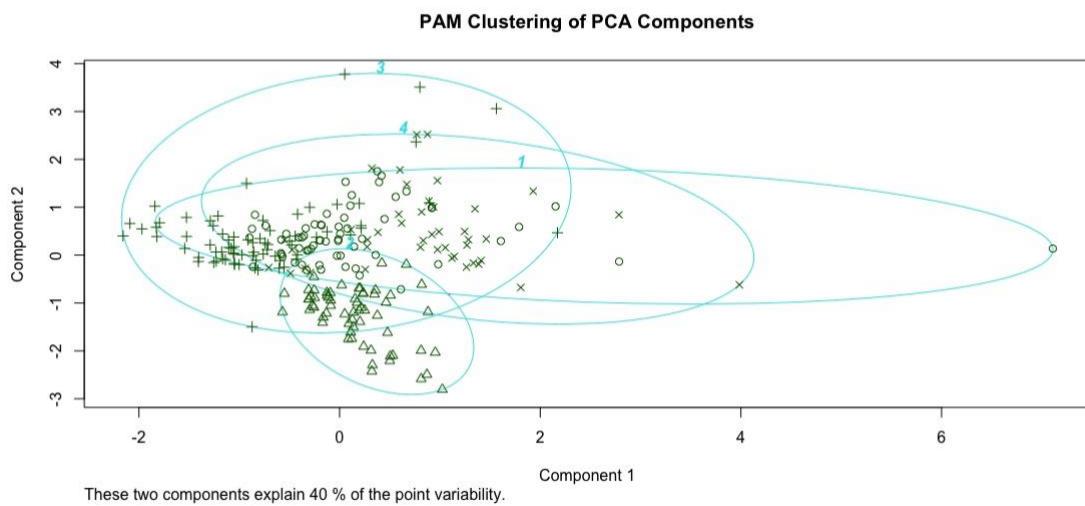


Figura 46 - PAM Clustering

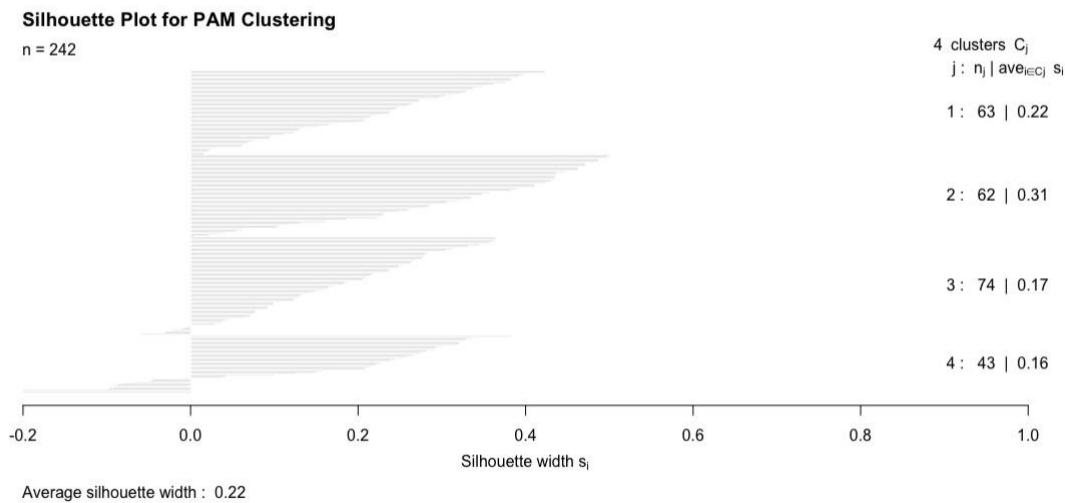


Figura 47 - Silhouette Plot PAM Clustering

O método PAM (Partitioning Around Medoids), é uma variação do K-Means que utiliza os Medóides para formar os clusters, ao contrário da média usada pelo K-Means. Por isso, em teoria, é um algoritmo que lida melhor com outliers, por trabalhar com valores absolutos e não médias, mas para esse problema obteve um desempenho pior no Silhouette Plot em comparação ao K-Means. O resultado acima foi obtido utilizando a métrica de distância Euclidiana (em linha reta), também foi testado um modelo com a métrica Manhattan, mas esse obteve um resultado ainda pior.

Gaussian Mixture Model:

O GMM é um método probabilístico que assume que os dados podem ser modelados como uma combinação de múltiplas distribuições normais (Gaussianas). Cada cluster é representado por uma Gaussiana, definida por parâmetros como média e variância.

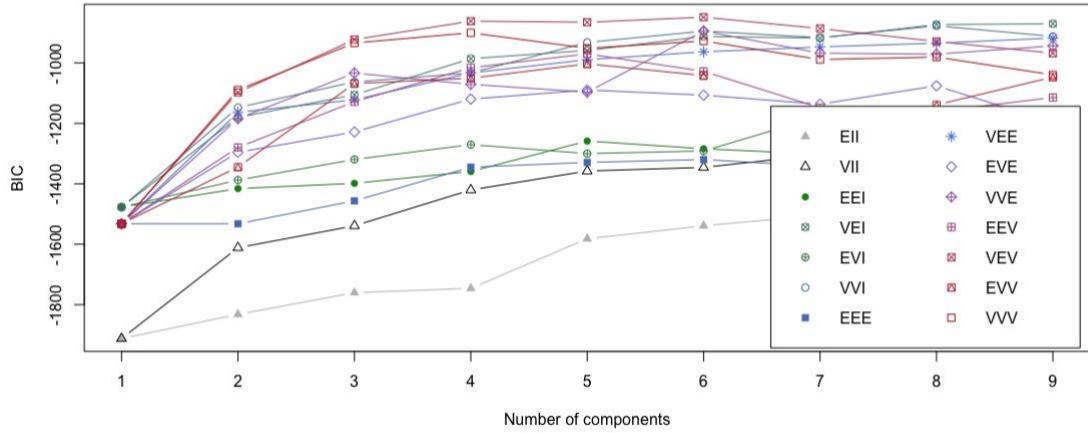


Figura 48 – Seleção de modelo do GMM

Como o gráfico acima demonstra, o modelo selecionado para determinar as formas dos *clusters* foi o *VEV* (Variância Elipsoidal), o que significa que os clusters terão formatos elípticos, alongados, e não apenas circulares. Para selecionar o número de clusters foi utilizado o método *BIC* (*Bayesian Information Criterion*), que busca encontrar o número ótimo de clusters e penaliza a complexidade excessiva do modelo. Nesse caso o número de *clusters* escolhido foi 6.

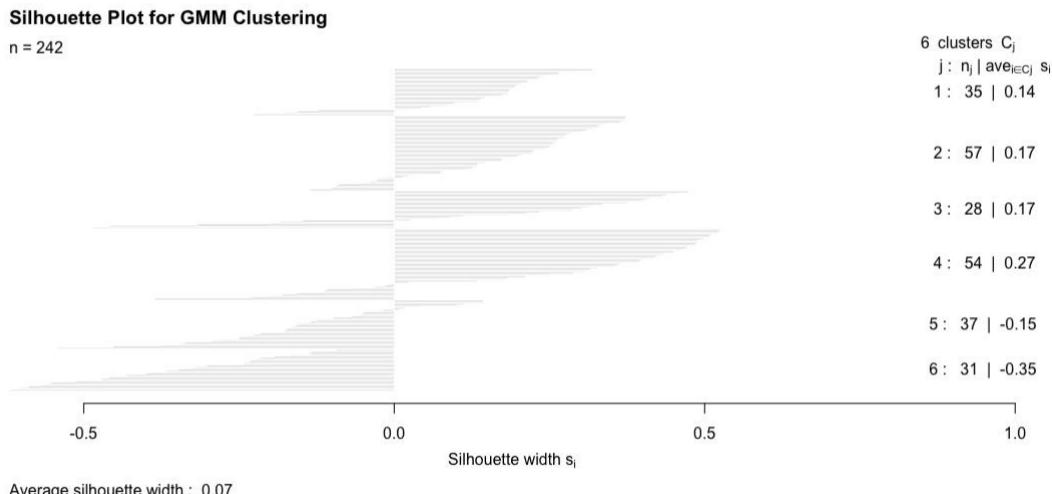


Figura 49 - Silhouette Plot GMM Clustering

Ao executar o algoritmo ele obteve um resultado muito insatisfatório, o que sugere que os *clusters* não estão claramente separados. Dos clusters sugeridos, nenhum obteve uma pontuação maior do que 0,27 na escala silhueta, e dois obtiveram pontuações negativas, o que indica que as regiões dentro desses *clusters* podem estar mal atribuídas.

Variáveis de *Profile* em comparação aos *Clusters* formados pelo *K-Means*:

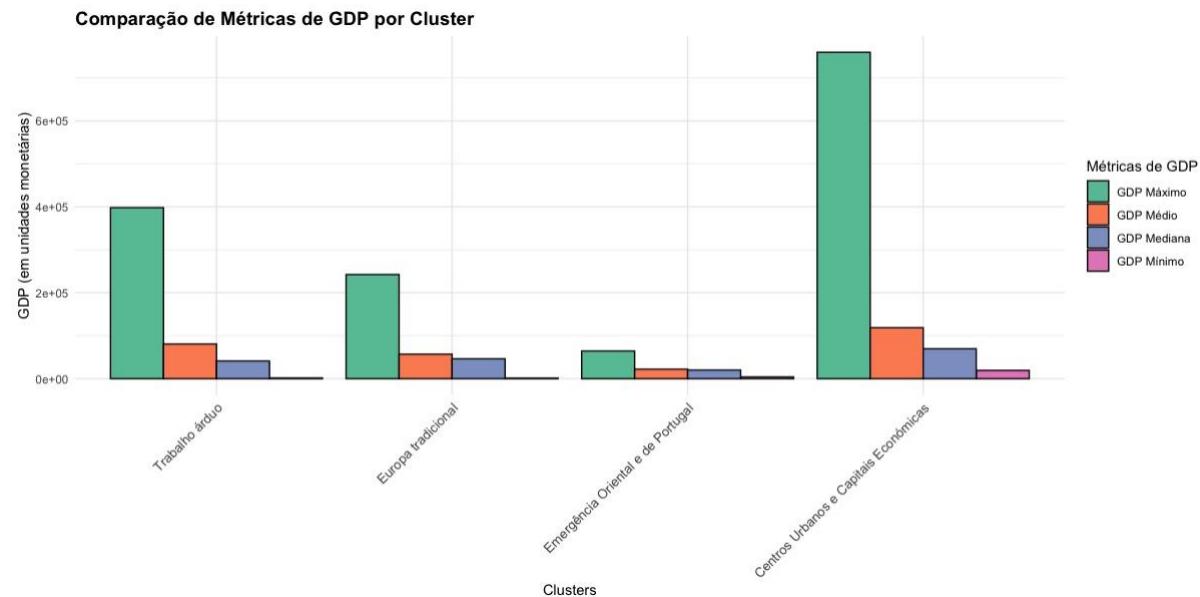


Figura 50 - GDP por Cluster

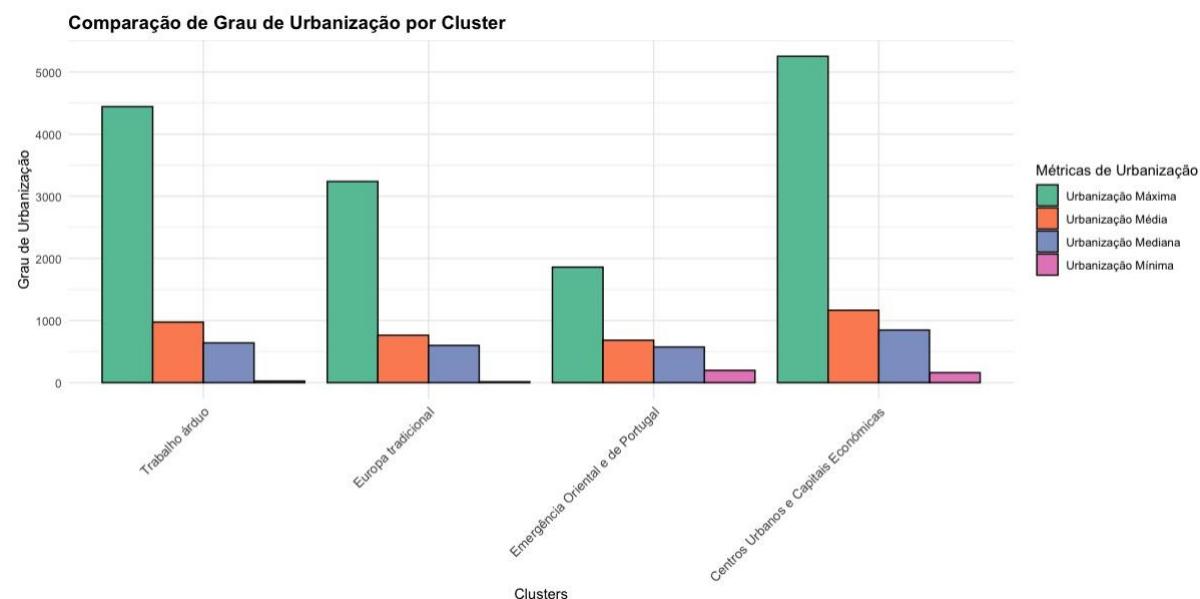


Figura 51 - Grau de Urbanização por Cluster

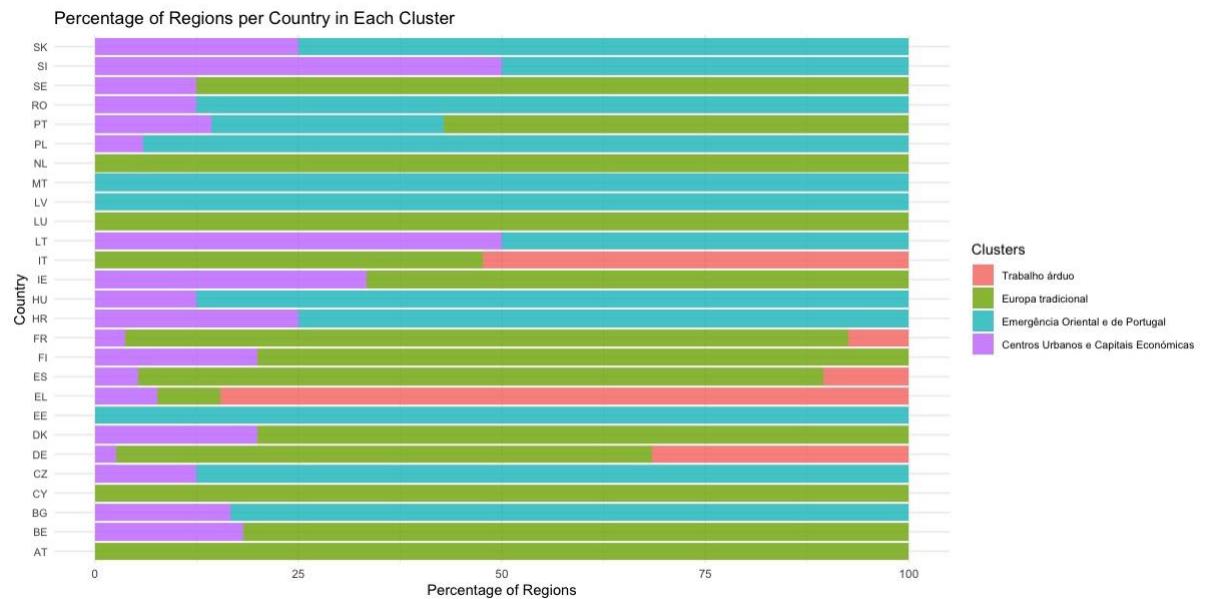


Figura 52 - Percentagem de Regiões nos Clusters

Distribuição de Capitais por Cluster

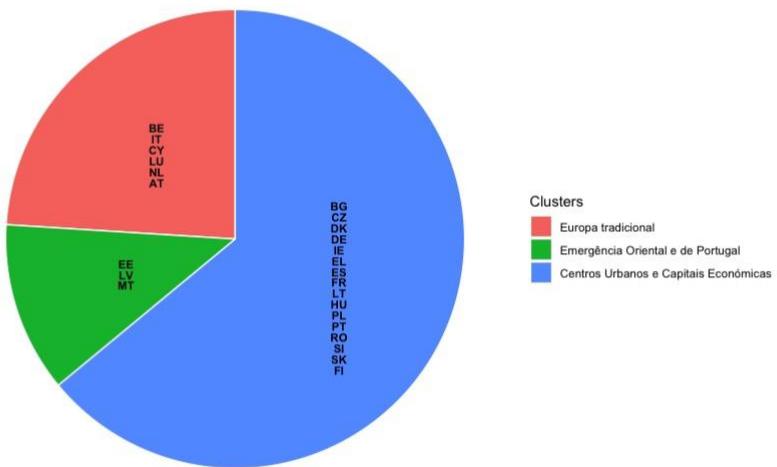


Figura 53 - Distribuição de Capitais por Cluster

O mistério da Andaluzia

Apesar do Elbow Plot (Figura 44) indicar um número ideal de 4 *clusters*, é possível, pela interpretação do mesmo, argumentar que 6 *clusters* pode ser um bom número também. Ao explorar essa possibilidade nos deparamos com uma situação interessante, um *cluster* com uma só região, a ES61 (Andaluzia).

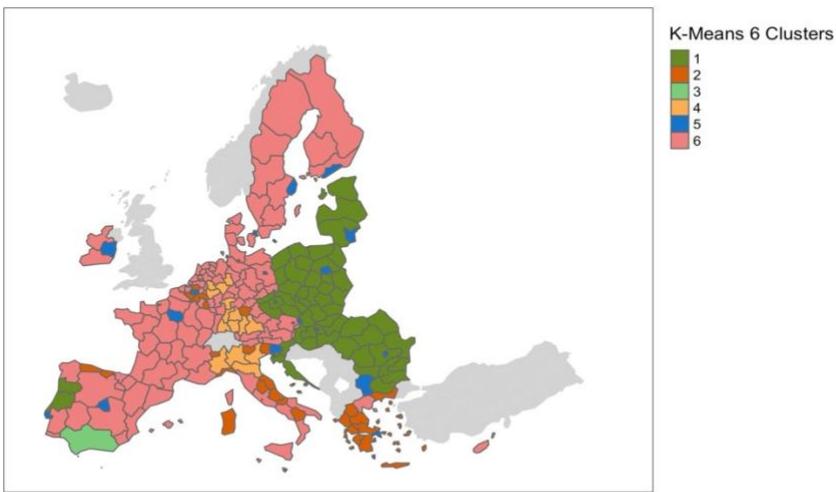


Figura 54 - Mapa K-Means com 6 Clusters

Os restantes *clusters* são quase idênticos aos encontrados pelo modelo com 4 *clusters*, que foi o escolhido para nortear as conclusões do relatório. O *cluster* a mais, além do ES61, é formada pelas ricas regiões industriais da Baviera e do Norte da Itália, enquanto as regiões da Grécia, que estavam juntas com as zonas Ítalo-Germânicas no *cluster* “Trabalho Árduo”, ganharam a companhia de regiões mais empobrecidas do *cluster* “Europa tradicional”.

Mesmo ao tentar uma separação com mais *clusters*, ignorando a orientação do *Elbow Plot*, a região ES61 continuava sozinha no seu próprio *cluster*.

Após extensa análise, os dados revelaram que a situação do trabalho na região da Andaluzia em 2019 realmente era um caso distinto. É a região NUTS 2 que concentra a maior população na Espanha e a terceira entre as regiões analisadas, ficando atrás apenas da FR10 (Ille de France) e ITC4 (Lombardia). Essa concentração populacional, em uma das principais economias europeias, faz com que haja um grande volume de trabalho em setores como a administração pública, o varejo e construção civil, elementos importantes do componente “Dimensão do Emprego Geral”. Mas, ao contrário dessas outras duas regiões, a Andaluzia apresenta uma baixa produtividade do trabalho, com muitas pessoas empregadas em setores de menor valor agregado, como agricultura e turismo, o que demonstra que a região sofre com grandes desigualdades internas, fatores que afetam os componentes “Emprego Industrial” e “Inovação vs Tradição”. Por último, outro fator importante para o destaque da Andaluzia foram as altas taxas de desemprego estrutural e de longo prazo, componente

“Raízes do Desemprego”, que afetava a região no ano analisado, especialmente se formos comparar com grandes regiões de importância interna similar.