VALIDAÇÃO DE MODELOS DE CLUSTERIZAÇÃO

Shirlei Dalila Machado Rezende

https://github.com/dalilamachado/clusterizacao

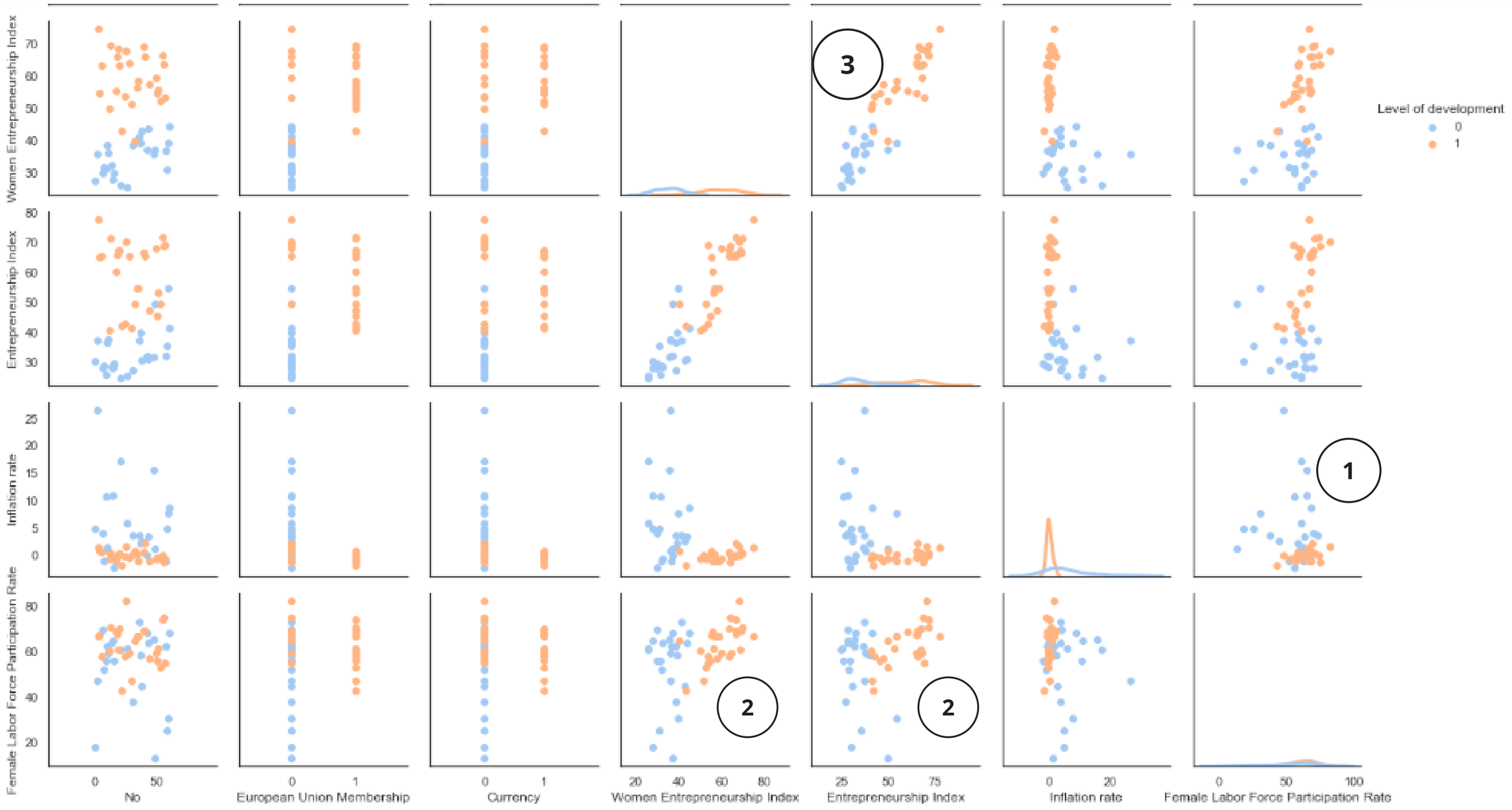
BASE DE DADOS

Base de dados: Women Entrepreneurship and Labor Force

Justificativa: No Sebrae Nacional trabalhamos com iniciativas voltadas para o empreendedorismo feminimo. A base escolhida no Kaggle ajuda a entender a posição do Brasil em relação aos outros países.

Objetivos:

1. Comparar o empreendedorismo feminino do Brasil com outros países.


Level of development
0
1
No
European Union Membership
Currency
Women Entrepreneurship Index
Entrepreneurship Index
Inflation rate
Female Labor Force Participation Rate
3
1
2
2

ANÁLISE FAIXA DINÂMICA

1. A inflação é mais baixa nos países onde há maior concentração de empreendedorismo feminino

2. A relação força de trabalho feminina x empreendedorismo feminina é bem equilibrada tanto em países desenvolvidos quanto em países em desenvolvimento.

3. A quantidade de empreendedoras femininas é maior em países desenvolvidos

PRÉ-PROCESSAMENTO

1. Verificar informações básicas do dataframe;
2. Transformar dados categóricos em numéricos das seguintes colunas: "Level of development", "European Union Membership", "Currency";
3. Excluir a coluna "No";
4. Retirar a coluna "Country" para fazer a análise. Ela foi adicionada novamente após a clusterização
5. Normalização dos dados

PERGUNTAS TEÓRICAS

Com os resultados em mão, descreva o processo de mensuração do índice de silhueta. Mostre o gráfico e justifique o número de clusters escolhidos.

A silhueta interpreta a consistência do conjunto de dados. Subtrai a distância da amostra até o ponto do outro cluster, subtrai a distância da amostra até o centroide e divide pelo máximo entre os dois.
Interpretação:
Separação ideal entre os clusters: 1
Separação confusa: 0
Separação ruim: - 1

Para a base de dados escolhida, a melhor silhueta para o Kmeans foi de 0.53, o que representa atenção à classificação realizada. O principal dado usado para separar os clusters, foi a taxa de inflação, seguido pela taxa de empreendedorismo feminino e taxa geral de empreendedorismo.

A silhueta é um o índice indicado para escolher o número de clusters para o algoritmo de DBScan?

Não. A silhueta funciona bem em clusters convexos, portanto é um indice indicado para o K-means. Como o DBSCAN funciona melhor para clusters não convexos, a silhueta pode apresentar resultado inferior para o DBSCAN, caso ele seja comparado com outro modelo.

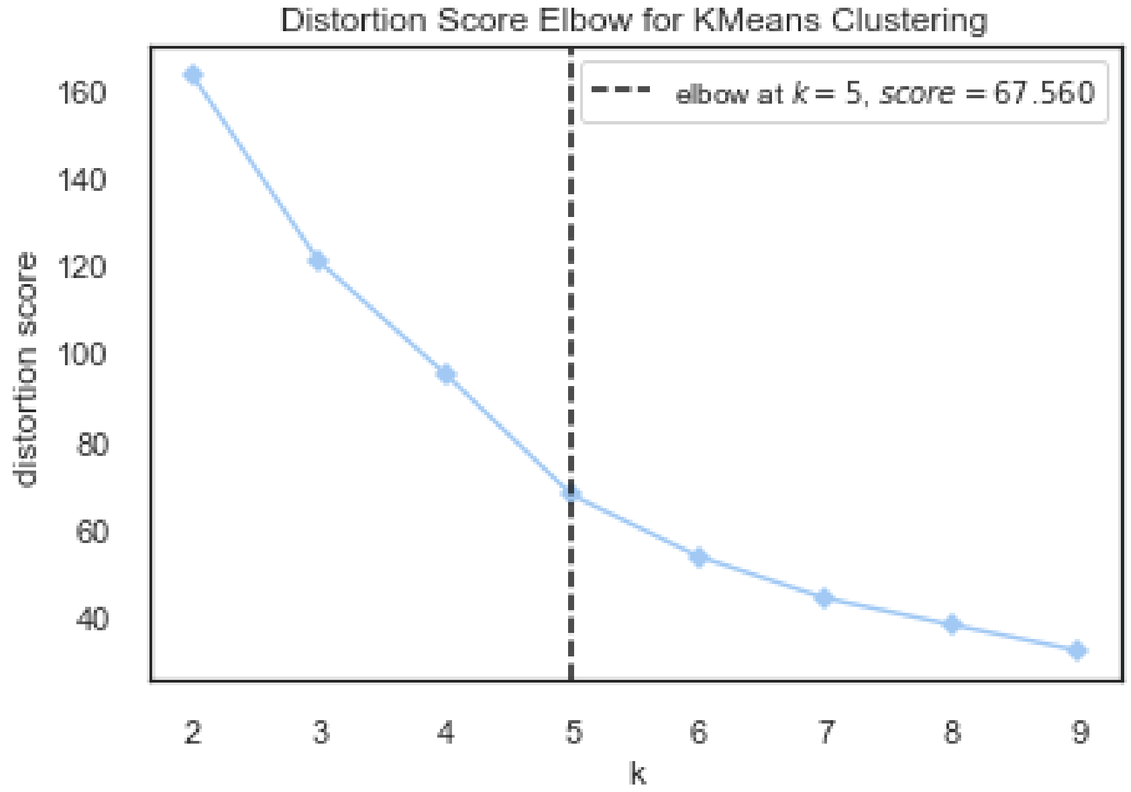
K-MEANS

Escolha de clusters

Método silhueta

```
Silhouette Score for k = 2: 0.477
Silhouette Score for k = 3: 0.453
Silhouette Score for k = 4: 0.428
Silhouette Score for k = 5: 0.447
```

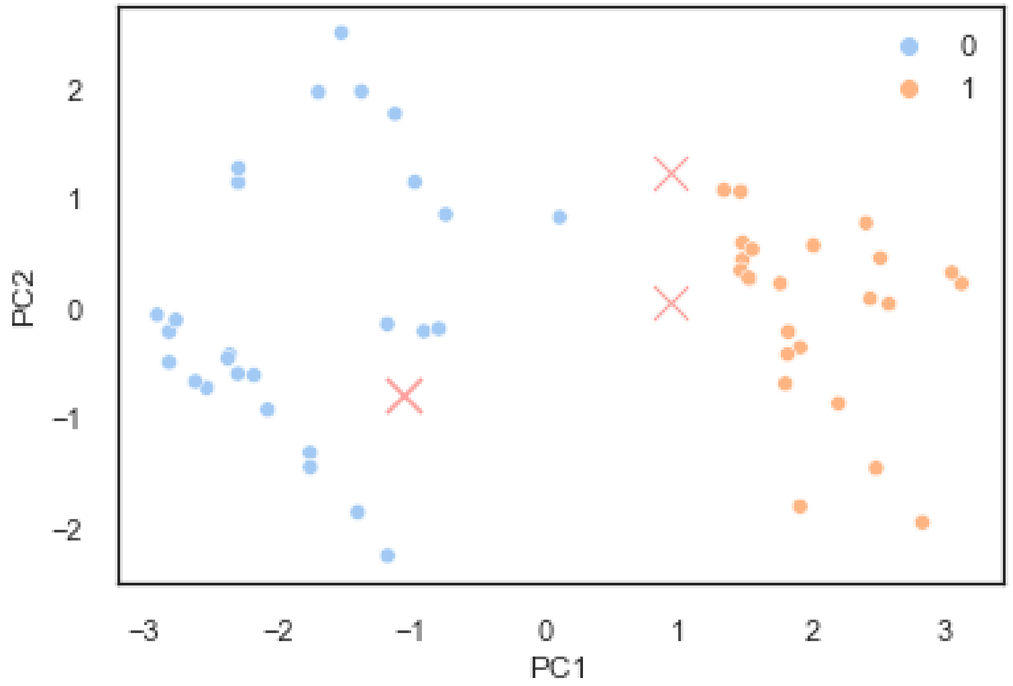
Método cotovelo



K-MEANS

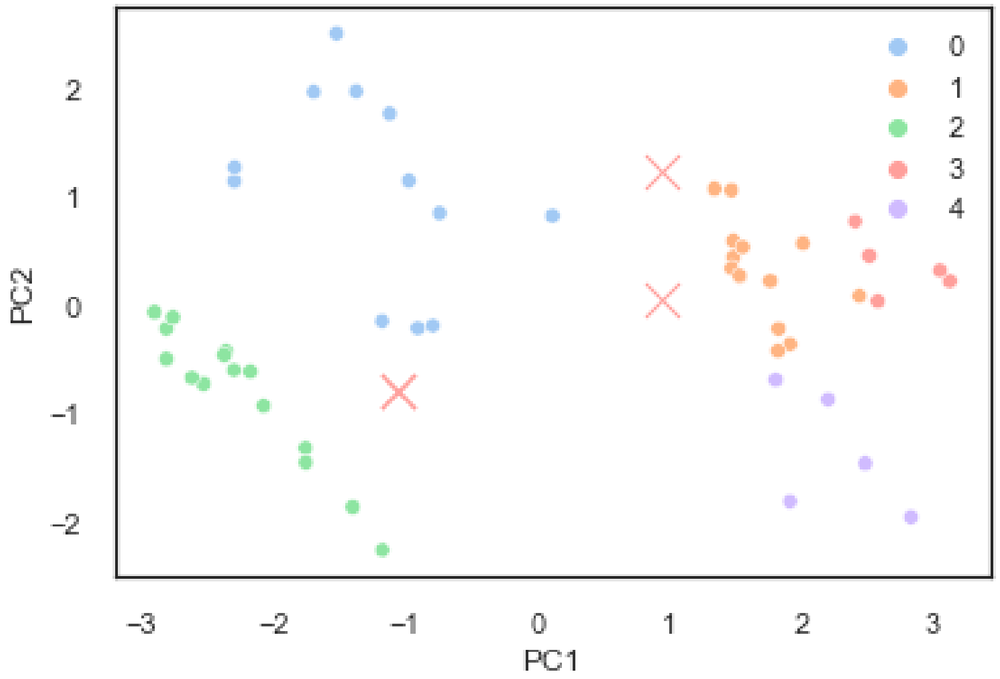
Comparação

Resultado K-mens - silhueta



Sihouette score para 2 clusters: 0.43

Resultado K-mens - cotovelo



Sihouette score para 5 clusters: 0.52

O K-means apresentou melhor score para o método cotovelo, com 5 clusters

DBSCAN

O que deu errado

Tentei aplicar o DBCV como medida de validação, mas não foi possível.
O código apresentou 2 erros:

- **key = 0**
- **nan**

Para o erro key = 0 encontrei uma solução explicando que deveria transformar o dataset em um array numpy. Fiz isso, mas retornou como erro nan, que não consegui corrigir.

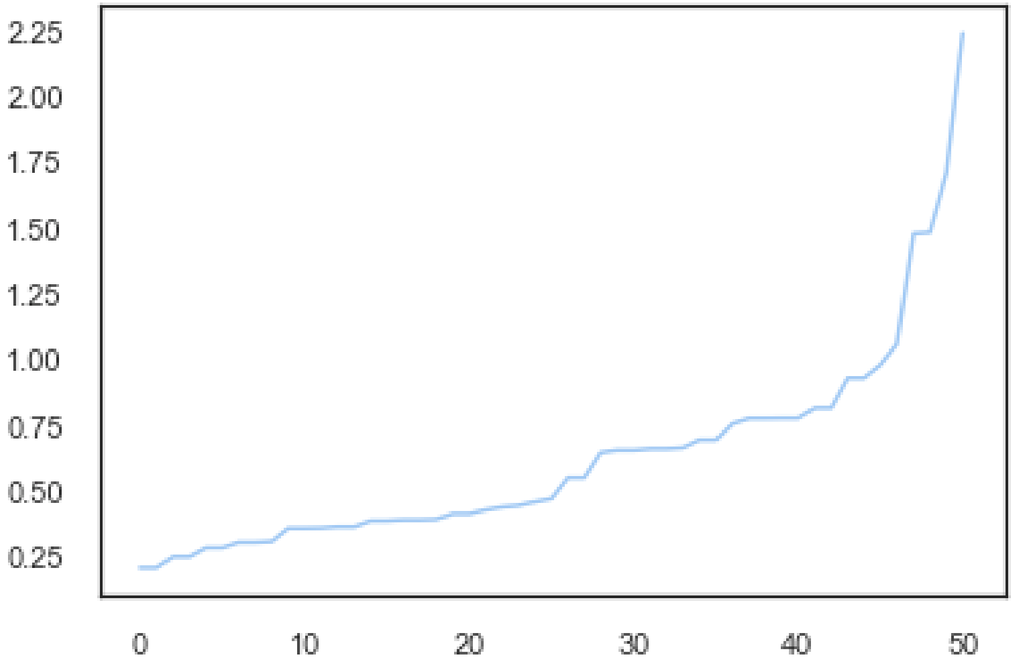
Encontrei outras formas de validação aqui: https://bit.ly/3W8G4uN, mas pesquisando descobri que são para clusters convexos, ou seja, não servem para o DBSCAN.

Portanto, a medida de validação usada no trabalho para o DBSCAN foi apenas a silhueta, apesar de não ser a medida ideal.

DBSCAN

Escolha de clusters

Silhueta para identificação de quantidade de eps e min_samples

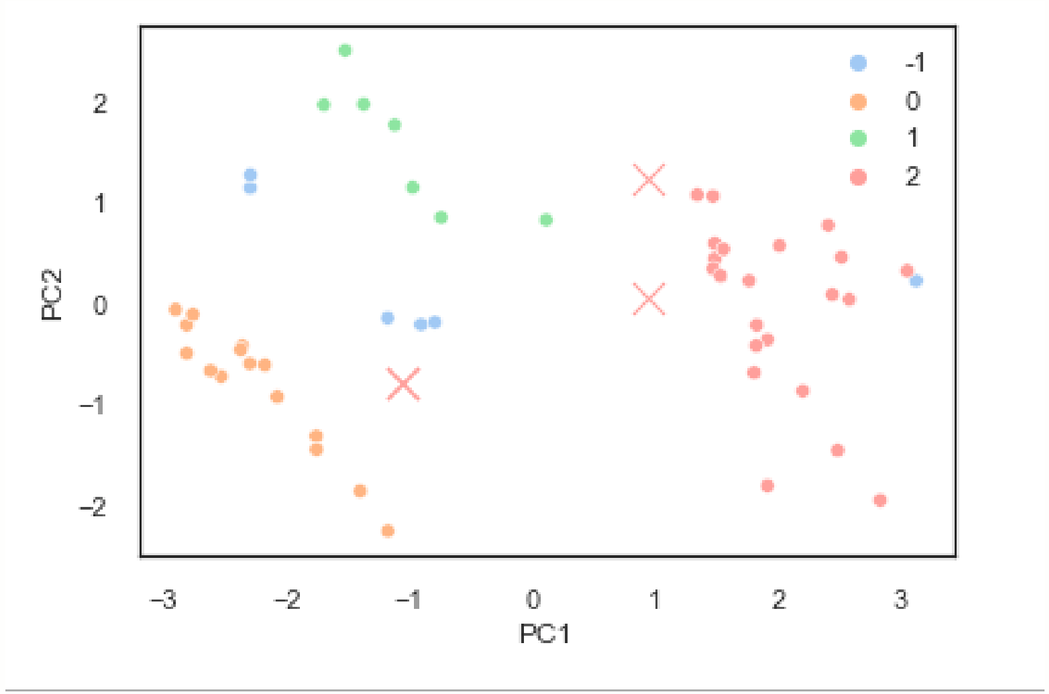


Eps: 1.9
min_samples: 6

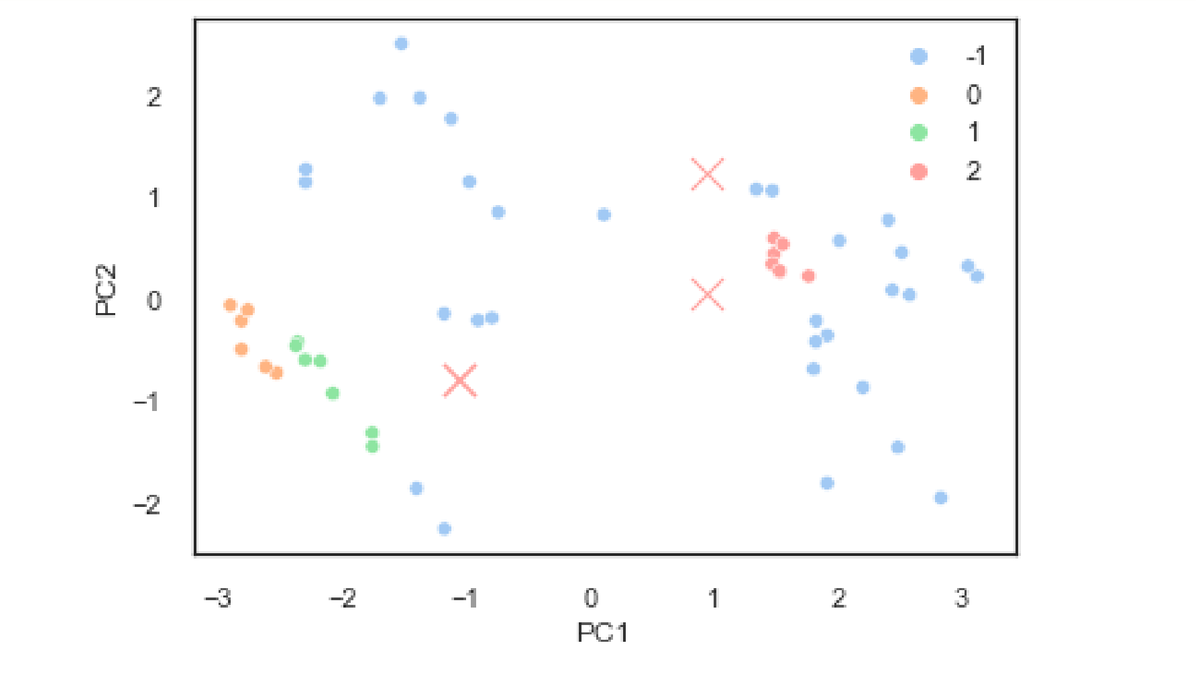
Método: nearest neighbors
Eps: 1.2
min_samples: 5

DBSCAN

Comparação



Eps: 1.9
min_samples: 6
Silhouette score: 0.12



Eps: 1.2
min_samples: 5
Silhouette score: 0.54

A escolha de eps e min_samples pelo nearest neighbors gerou um score bastante superior se comparado com o método silhueta

K-MEANS

Interpretação

Cluster 0: A inflação é variada e o agrupamento foi realizado pelos altos indices de empreendedorismo feminino

Country	Australia	Croatia	Denmark	Hungary	Iceland	Japan	Norway	Poland	Singapore	Sweden	Switzerland	Taiwan	Totals
Women Entrepreneurship Index	74.8	49.9	69.7	53.7	68.0	40.0	66.3	57.7	59.8	66.7	63.7	53.4	
Entrepreneurship Index	77.6	40.6	71.4	42.7	70.4	49.5	65.6	47.4	68.1	71.8	68.6	69.1	
Inflation rate	1.5	-0.5	0.5	-0.1	1.6	0.8	2.17	-0.9	-0.5	0.0	-1.1	-0.61	
Totals	10.3%	6.9%	9.6%	7.4%	9.4%	5.5%	9.2%	8.0%	8.3%	9.2%	8.8%	7.4%	100.0%

Cluster 1: Baixa inflação e baixos indices de empreendedorismo

Country	Bolivia	Bosnia and Herzegovina	China	Costa Rica	Ecuador	El Salvador	India	Macedonia	Malaysia	Mexico	Panama	Peru	Thailand	Uruguay	Totals
Women Entrepreneurship Index	29.7	31.6	38.3	36.1	32.3	29.9	25.3	41.2	39.2	42.8	36.9	43.6	36.6	44.5	
Entrepreneurship Index	28.0	28.9	36.4	37.7	28.2	29.6	25.3	37.1	40.0	30.7	32.2	30.9	32.1	41.4	
Inflation rate	4.1	-1.0	1.4	0.8	-0.5	-2.25	5.9	3.7	2.3	2.7	0.1	3.5	-0.9	8.67	
Totals	5.8%	6.2%	7.5%	7.1%	6.4%	5.9%	5.0%	8.1%	7.7%	8.4%	7.3%	8.6%	7.2%	8.8%	100.0%

K-MEANS

Interpretação

Cluster 2: Inflação baixa e altos indices de empreendedorismo

Country	Austria	Belgium	Estonia	Finland	France	Germany	Greece	Ireland	Italy	Latvia	Lithuania	Netherlands	Slovakia	Slovenia	Spain
Women Entrepreneurship Index	54.9	63.6	55.4	66.4	68.8	63.6	43.0	64.3	51.4	56.6	58.5	69.3	54.8	55.9	52.5
Entrepreneurship Index	64.9	65.5	60.2	65.7	67.3	67.4	42.0	65.3	41.3	54.5	54.6	66.5	45.4	53.1	49.6
Inflation rate	0.9	0.6	-0.88	-0.2	0.0	0.5	-1.7	-0.3	0.0	0.2	-0.9	0.6	-0.3	-0.5	-0.5
Totals	6.2%	7.2%	6.3%	7.6%	7.8%	7.2%	4.9%	7.3%	5.8%	6.4%	6.7%	7.9%	6.2%	6.4%	6.0%

Cluster 3: Alta inflação e baixos indices de empreendedorismo

Country	Argentina	Brazil	Egypt	Ghana	Russia	Totals
Women Entrepreneurship Index	35.7	31.1	27.7	25.8	35.6	
Entrepreneurship Index	37.2	25.8	28.1	24.8	31.7	
Inflation rate	26.5	10.67	11.0	17.2	15.5	
Totals	22.9%	19.9%	17.8%	16.5%	22.8%	100.0%

K-MEANS

Interpretação

Cluster 4: Inflação variada e baixos indices de empreendedorismo

Country	Algeria	Jamaica	Saudi Arabia	Tunisia	Turkey	Totals
Women Entrepreneurship Index	27.4	38.6	37.0	30.7	39.3	
Entrepreneurship Index	30.2	27.2	49.6	35.5	54.6	
Inflation rate	4.8	3.7	1.2	4.8	7.7	
Totals	15.8%	22.3%	21.4%	17.7%	22.7%	100.0%

DBSCAN

Interpretação

Cluster -1: o agrupamento é ruim e não mostra um padrão claro.

Country	Algeria	Argentina	Brazil	Croatia	Denmark	Egypt	Ghana	Hungary	Jamaica	Japan	Poland	Russia	Saudi Arabia	Sweden	Taiwan	Tunisia	Turkey	Totals
Women Entrepreneurship Index	27.4	35.7	31.1	49.9	69.7	27.7	25.8	53.7	38.6	40.0	57.7	35.6	37.0	66.7	53.4	30.7	39.3	
Entrepreneurship Index	30.2	37.2	25.8	40.6	71.4	28.1	24.8	42.7	27.2	49.5	47.4	31.7	49.6	71.8	69.1	35.5	54.6	
Inflation rate	4.8	26.5	10.67	-0.5	0.5	11.0	17.2	-0.1	3.7	0.8	-0.9	15.5	1.2	0.0	-0.61	4.8	7.7	
Totals	3.8%	5.0%	4.3%	6.9%	9.7%	3.8%	3.6%	7.5%	5.4%	5.6%	8.0%	4.9%	5.1%	9.3%	7.4%	4.3%	5.5%	100.0%

Cluster 0: agrupa por países com infação próxima a 0 ou menor que 0.

Country	Austria	Belgium	Estonia	Finland	France	Germany	Greece	Ireland	Italy	Latvia	Lithuania	Netherlands	Slovakia	Slovenia	Spain	Totals
Women Entrepreneurship Index	54.9	63.6	55.4	66.4	68.8	63.6	43.0	64.3	51.4	56.6	58.5	69.3	54.8	55.9	52.5	
Entrepreneurship Index	64.9	65.5	60.2	65.7	67.3	67.4	42.0	65.3	41.3	54.5	54.6	66.5	45.4	53.1	49.6	
Inflation rate	0.9	0.6	-0.88	-0.2	0.0	0.5	-1.7	-0.3	0.0	0.2	-0.9	0.6	-0.3	-0.5	-0.5	
Totals	6.2%	7.2%	6.3%	7.6%	7.8%	7.2%	4.9%	7.3%	5.8%	6.4%	6.7%	7.9%	6.2%	6.4%	6.0%	100.0%

DBSCAN

Interpretação

Cluster 1: agrupa por países com inflação entre -0.5 e 2.17

Country	Australia	Iceland	Norway	Singapore	Switzerland	Totals
Women Entrepreneurship Index	74.8	68.0	66.3	59.8	63.7	
Entrepreneurship Index	77.6	70.4	65.6	68.1	68.6	
Inflation rate	1.5	1.6	2.17	-0.5	-1.1	
Totals	22.5%	20.4%	19.9%	18.0%	19.2%	100.0%

Cluster 2: as taxas de inflação são bastante variadas, mas os indices de empreendedorismo são mais baixos

Country	Bolivia	Bosnia and Herzegovina	China	Costa Rica	Ecuador	El Salvador	India	Macedonia	Malaysia	Mexico	Panama	Peru	Thailand	Uruguay	Totals
Women Entrepreneurship Index	29.7	31.6	38.3	36.1	32.3	29.9	25.3	41.2	39.2	42.8	36.9	43.6	36.6	44.5	
Entrepreneurship Index	28.0	28.9	36.4	37.7	28.2	29.6	25.3	37.1	40.0	30.7	32.2	30.9	32.1	41.4	
Inflation rate	4.1	-1.0	1.4	0.8	-0.5	-2.25	5.9	3.7	2.3	2.7	0.1	3.5	-0.9	8.67	
Totals	5.8%	6.2%	7.5%	7.1%	6.4%	5.9%	5.0%	8.1%	7.7%	8.4%	7.3%	8.6%	7.2%	8.8%	100.0%

PERGUNTAS TEÓRICAS

Um determinado problema, apresenta 10 séries temporais distintas. Gostaríamos de agrupá-las em 3 grupos, de acordo com um critério de similaridade, baseado no valor máximo de correlação cruzada entre elas. Descreva em tópicos todos os passos necessários.

1. Identificar o objetivo da similaridade. Exemplo: tendência
2. Monta-se a matriz de distância para criar a relação cruzada
3. Calcula-se a relação cruzada. Exemplo: item 1 do dataset correlacionado com todo o comprimento do item 2 - percorre-se de um ponto ao outro do eixo temporal.
4. A correlação máxima será o pico entre os dois itens.

Para o problema da questão anterior, indique qual algoritmo de clusterização você usaria. Justifique

Clusterização hierárquica, pois trabalha de forma equivalente a datasets com features e também com matriz de distância.

Indique um caso de uso para essa solução projetada.

Análise de tendência para mercado financeiro.

Sugira outra estratégia para medir a similaridade entre séries temporais. Descreva em tópicos os passos necessários.

DTW - Dynamic time warpin, que calcula alinhamentos que minimiza distâncias entre as séries.
Passos necessários:

1. Inserir ou remover entradas duplicadas adjacentes com custo zero
2. Editar uma posição a para b com custo da distânciad(b,a)
3. Inserir ou remover zero no final de um vetor com custo zero.

REFERÊNCIAS

https://github.com/IDB-FOR-DATASCIENCE/Unsupervised-ML-Modelling-for-Segmentation/blob/main/Segmentation%20Notebo

https://www.delftstack.com/howto/python-pandas/how-to-convert-pandas-dataframe-to-numpy-array/

https://medium.com/pizzadedados/kmeans-e-metodo-do-cotovelo-94ded9fdf3a9

https://larissaakemi.medium.com/testando-k-means-e-dbscan-investsp-prospect-advisor-139c57c47f59

https://www.kaggle.com/code/umairaslam/dbscan-clustering

https://medium.com/towards-data-science/machine-learning-clustering-dbscan-determine-the-optimal-value-for-epsilon-eps-p
3100091cfbc

https://www.kaggle.com/datasets/babyoda/women-entrepreneurship-and-labor-force