Projet analyse de données

Lucas Mauge

L'analyse des données socio-économiques est nécessaire à la compréhension des tendances et des disparités dans le développement des pays membres de l'Organisation de Coopération et de Développement Économiques (OCDE). Dans le cadre d'un projet d'analyse de données, nous nous pencherons sur un ensemble de données provenant de l'Observatoire de l'OCDE, couvrant plusieurs variables telles que le taux de natalité, le chômage, le PIB par habitant, et bien d'autres, pour différentes années et pays membres de l'OCDE. L'analyse que nous allons fournir a pour but d'explorer les relations qu'il peut y avoir entre certaines variables et ainsi découvrir des tendances significatives qui pourraient aider à comprendre les dynamiques socio-économiques au sein des pays de l'OCDE sur 4 années (1975 1977 1979 1981).

Nous commencerons par une analyse univariée qui permettra d'examiner certaines de ces variable individuellement pour obtenir un aperçu de leur distribution et de leurs caractéristiques statistiques. Nous poursuivrons par une analyse bivariée où nous explorerons les relations entre les paires de variables pour identifier d'éventuelles corrélations. La Troisième analyse sera une analyse en composantes principales (ACP). En effet ous utiliserons l'ACP pour réduire le nombres de variables et visualiser les relations entre les observations et les variables. Enfin, avant de conclure, nous feront une classification non supervisée pour regrouper les pays en fonction de leurs caractéristiques socio-économiques similaires.

Première partie: Analyse univariée

Nous allons analyser séparémment deux de ces variables qui sont le taux de chômage ainsi que le produit intérieur brut (PIB) par habitant car on peut supposer que l'un peut expliquer l'autre. Nous traçons deux tableaux montrant les moyennes et les variances pour les deux variables sur 4 ans. Le tableau du haut représente celles du taux de chômage tandis que celui du bas représente celles du PIB.

Pays	meanD	stD
AL	39.50	4.654747
AU	19.75	4.112988
BE	70.75	28.987066
CA	74.75	4.924429
DA	60.00	29.325757
ES	83.75	40.606855
EU	71.00	10.954451
FI	48.50	18.064698
FR	55.50	13.796135
IR	81.00	14.537308
IT	65.50	22.233608
JA	20.50	1.290994
NO	18.00	6.164414
PB	51.25	15.882380
PO	73.25	13.573872
RU	63.00	32.731229
SU	20.00	3.915780

Pays	moyenne	variance
AL	9701.50	2525.1315
AU	7333.00	1988.9954
BE	8765.25	2009.1019
CA	9225.25	1987.8574
DA	10085.00	2601.2113
ES	4067.00	1235.2312
EU	9764.25	2411.3479
FI	7734.50	2114.6223
FR	8690.50	2248.4397
IR	3666.25	1136.0118
IT	4590.75	1537.8109
JA	7211.50	2383.0727
NO	10289.25	3005.4676
PB	8527.75	2122.7765
PO	1922.00	387.4567
RU	6124.00	2310.7136
SU	11066.25	2471.9674

En analysant les deux tableaux on peut observer une tendance qui pourrait s'avérer significative; c'est que lorsque le PIB d'un pays est élevé alors son taux de chômage est plus faible que celui dont les PIB sont plus bas. Cela semble cohérent que les pays les plus riches ait un taux de chomâge plus faible que les pays pauvres.

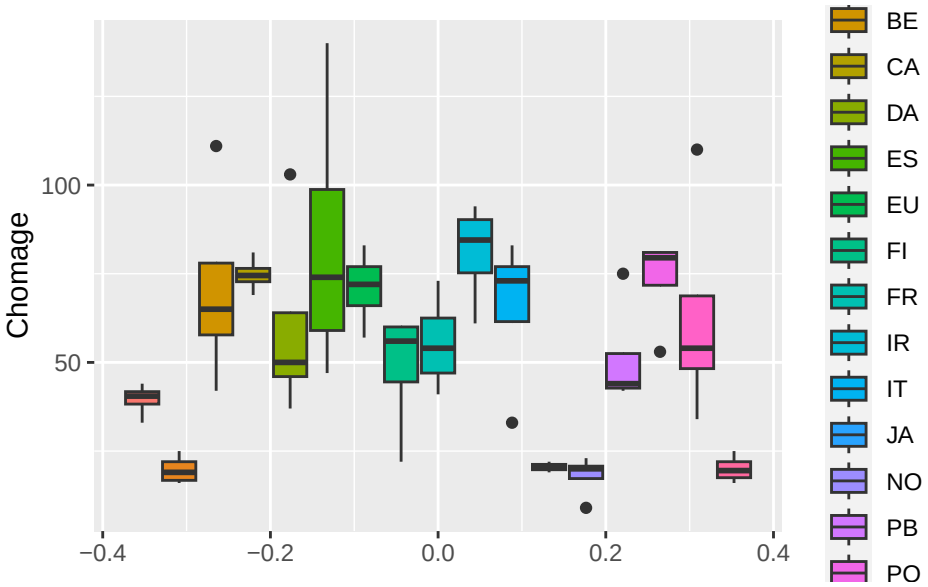
Pays	Natal	Chomage	a.prim	a.sec	pib	fbcf	infl	recc	m.inf	prot	nrj	Annee
AL	97.00	39.50	64.50	449.25	9701.50	216.75	49.00	429.25	156.25	64.50	427.25	78
AU	119.00	19.75	113.00	405.00	7333.00	259.50	63.00	431.25	161.75	58.00	336.50	78
BE	124.25	70.75	32.75	366.75	8765.25	205.25	78.50	424.25	133.25	62.75	454.25	78
CA	153.75	74.75	57.50	288.50	9225.25	232.75	91.25	370.75	130.50	65.50	898.25	78
DA	121.25	60.00	86.50	303.00	10085.00	201.00	105.25	473.00	91.50	69.00	375.50	78
ES	167.50	83.75	200.50	368.75	4067.00	209.50	168.25	266.50	132.75	52.50	189.75	78
EU	154.00	71.00	36.75	298.50	9764.25	174.50	84.25	322.50	140.00	73.00	814.75	78
FI	136.25	48.50	126.75	350.25	7734.50	252.00	118.50	403.75	93.50	65.50	515.25	78
FR	143.00	55.50	95.75	369.25	8690.50	221.25	106.00	431.25	109.75	74.00	346.75	78
IR	213.75	81.00	212.50	310.25	3666.25	273.50	149.25	371.00	150.50	67.75	244.25	78
IT	128.00	65.50	149.75	394.75	4590.75	200.00	160.50	367.50	169.50	52.00	242.50	78
JA	150.25	20.50	114.50	353.50	7211.50	308.50	86.50	251.25	85.00	35.50	309.75	78
NO	129.25	18.00	90.75	316.25	10289.25	316.25	91.50	515.00	94.25	62.50	536.25	78
PB	126.50	51.25	59.75	325.50	8527.75	207.75	70.50	548.25	92.00	62.25	454.25	78
PO	174.50	73.25	293.50	345.50	1922.00	213.00	198.75	269.25	354.25	38.75	99.50	78
RU	125.75	63.00	27.00	390.50	6124.00	179.25	142.50	400.00	138.75	55.00	370.75	78
SU	118.00	20.00	59.75	336.50	11066.25	200.00	99.00	580.50	76.50	70.00	609.50	78

On s'aperçoit que les pays nordiques tels que le Danemark, la Suède et la Norvège affichent un PIB élevé, tandis que des nations comme l'Espagne et le Portugal présentent un PIB plus bas ainsi qu'un taux de chômage, une inflation et un pourcentage d'actifs dans le secteur primaire élevés.On peut supposer qu'il existe des corrélations entre ces variables nous le verrons par la suite.

Annee	Natal	Chomage	a.prim	a.sec	pib	fbcf	infl	recc	m.inf	prot	nrj
75	147.7647	41.88235	116.76471	364.4706	5367.647	228.6471	108.2941	381.8824	155.7059	58.05882	401.4706
77	140.8235	51.88235	112.35294	352.4118	6490.588	232.2353	117.3529	405.4706	145.8824	60.17647	419.6471

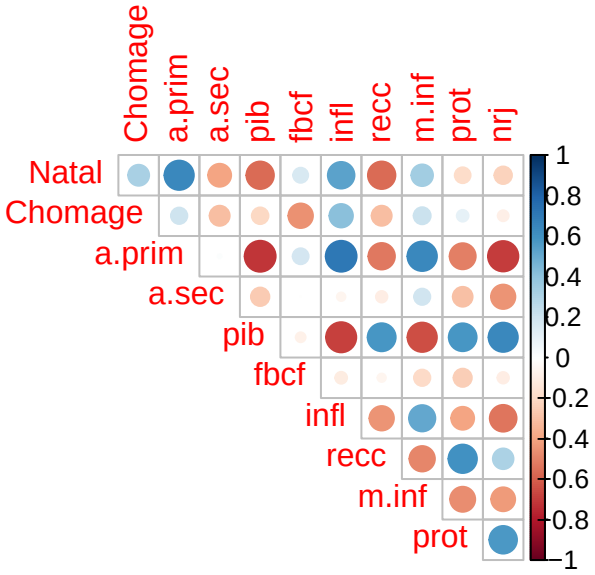
Annee	Natal	Chomage	a.prim	a.sec	pib	fbcf	infl	recc	m.inf	prot	nrj
79	137.4118	51.00000	104.11765	348.0588	9015.294	225.5294	106.9412	404.4118	130.9412	61.29412	452.7059
81	134.4706	70.76471	95.35294	340.2353	9423.882	224.3529	105.6471	421.2353	111.0000	62.47059	426.1765

Dans le tableau ci dessus on exprime les variable quantitative en fonction des années. On observe que plus les années avancent plus les PIB augmentent. L'augmentation du PIB entre 1975 et 1981 s'explique par la reprise économique après une période difficile (crise pétrolière de 1973), les progrès technologiques, la stabilité politique, les politiques de relance et l'expansion des échanges internationaux.



On peut donc observer que la variabilité du taux de chomage est très élevé en Espagne tandis que dans des pays comme la Norvège où il est bas la variabilité est très faible. Cela confirme bien ce que nous avons vu dans les tableaux précédents.

Partie 2:Analyse bivariée



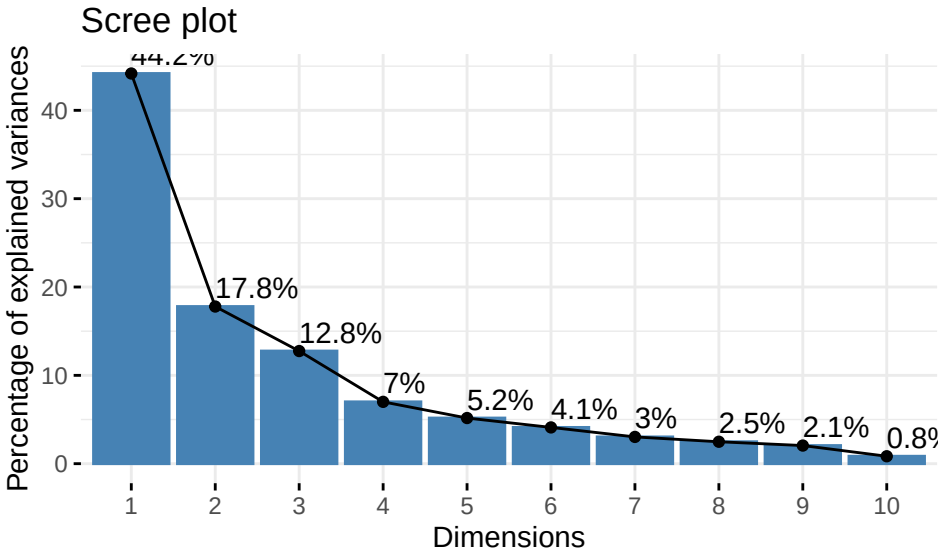
On observe que la taux de natalité est corrélé positivement au pourcentage d'actifs dans le secteur primaire et tous deux sont corrélés positivement à la mortalité infantile. Il est logique que la mortalité infantile grandisse lorsque le taux brute de natalité augmente. Le PIB est quand à lui corrélé positivement à la consommation d'énergie et à la consommation de protéines animales. Cela prend sens étant donné que plus un pays est riche, plus il peut consommer d'énergie et de protéines animales. Ces derniers sont fortement corrélés négativement au pourcentage d'actifs dans le secteur primaire ainsi qu'à la mortalité infantile.

Ce graphique de corrélation nous montre donc qu'il y a des variables qui représentent des pays plus développés que les autres et inversement. Cependant plusieurs de ces variables vont nous donner les mêmes informations comme l'inflation, la mortalité infantile et le pourcentage d'actifs dans le secteur primaire, c'est pourquoi nous allons procéder à une analyse en composantes principales.

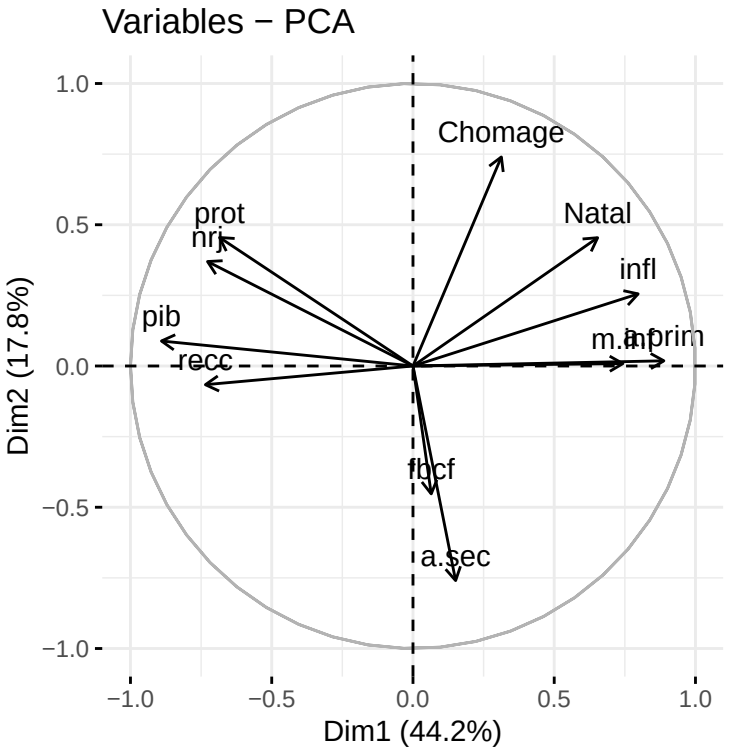
Trosième partié: Analyse en composante principales (ACP)

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	4.8583951	44.1672283	44.16723
comp 2	1.9567562	17.7886924	61.95592
comp 3	1.4026124	12.7510214	74.70694
comp 4	0.7700245	7.0002229	81.70717
comp 5	0.5684113	5.1673753	86.87454
comp 6	0.4516432	4.1058477	90.98039
comp 7	0.3335039	3.0318538	94.01224
comp 8	0.2731346	2.4830422	96.49528
comp 9	0.2257965	2.0526952	98.54798
comp 10	0.0927308	0.8430071	99.39099
comp 11	0.0669915	0.6090137	100.00000

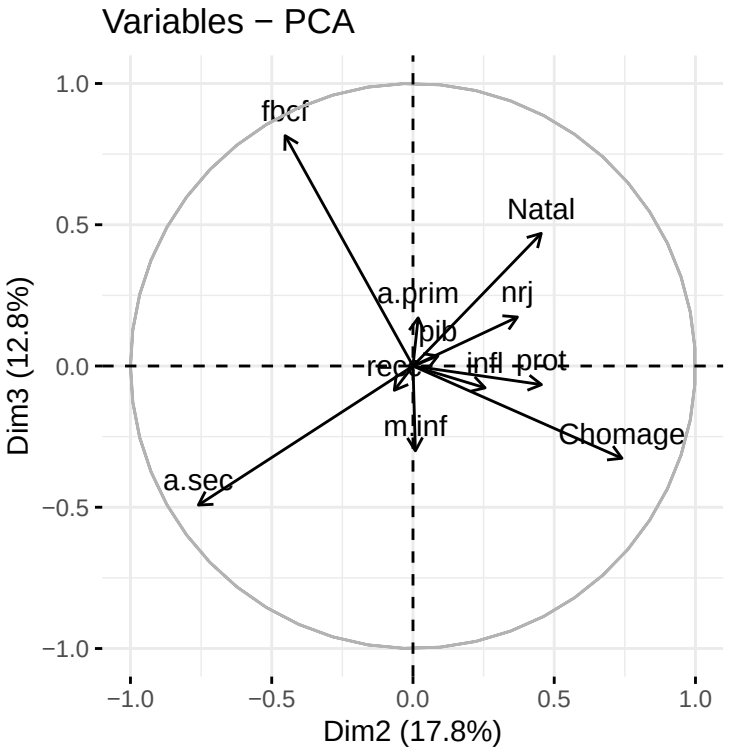
On observe que seul les trois premières composantes ont une eigenvalue supérieure à 1, on aurait donc tendance à garder ces trois composantes.



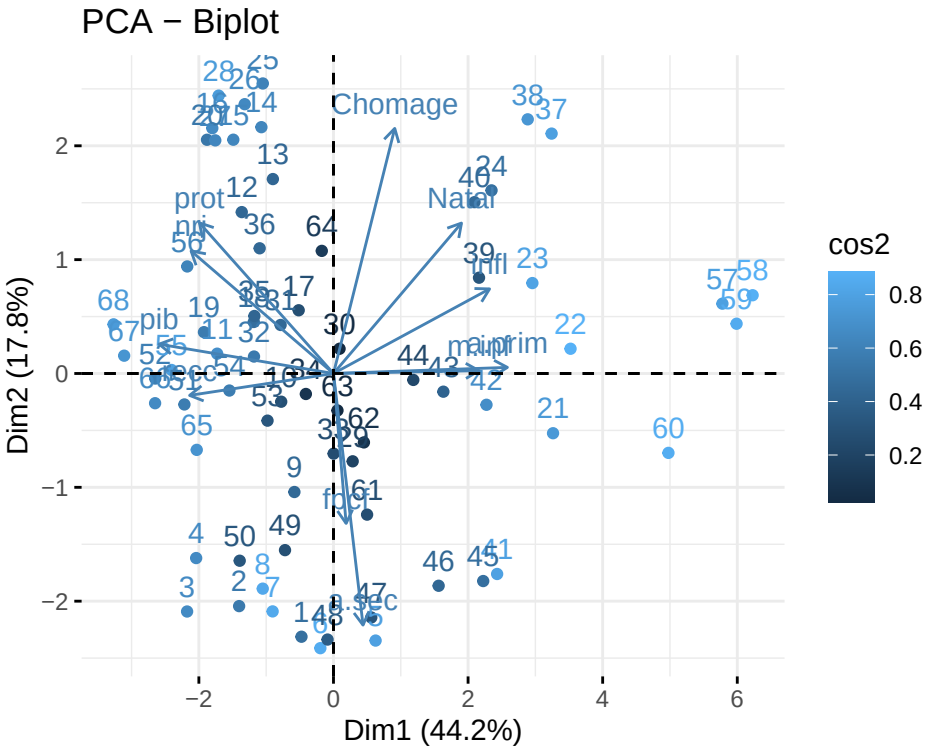
On a que la somme des eigenvalue est égale à l'intertie totale du nuage de points. Le pourcentage d'inertie porté par les trois axes principaux est d'environ 74%.



On a que toutes les flèches sont proches du cercle exceptée celle de la formation brute de capital fixe donc cela signifie que sur ce plan les variables sont toutes bien expliquées en dehors de celle désignant la formation brute de capital fixe et pourront donc être utilisées pour l'interprétation des deux axes. Les flèches associées aux variables chomage, natal, infl, a.prim sont proches et vont dans le même sens: ces variables sont toutes corrélées et positivement. De plus elles sont corrélées aux flèches des variables prot nrj pib et recc négativement. On observe un angle d'environ 90 entre ces variables et la variable a.sec signifiant que ces deux groupes de variables sont peu corrélés. L'axe 1 de notre analyse semble représenter une mesure globale des aspects socio-économiques liés à la consommation de protéines animales, d'énergie, de PIB et de recettes courantes par habitant, opposant ainsi les pays affichant des niveaux élevés dans ces domaines à ceux présentant des niveaux plus faibles. En revanche, l'axe 2 pourrait être associé au pourcentage d'actifs travaillant dans le secteur secondaire (a.sec), distinguant ainsi les pays où une part importante de la population active est employée dans ce secteur de ceux où cette proportion est moindre.

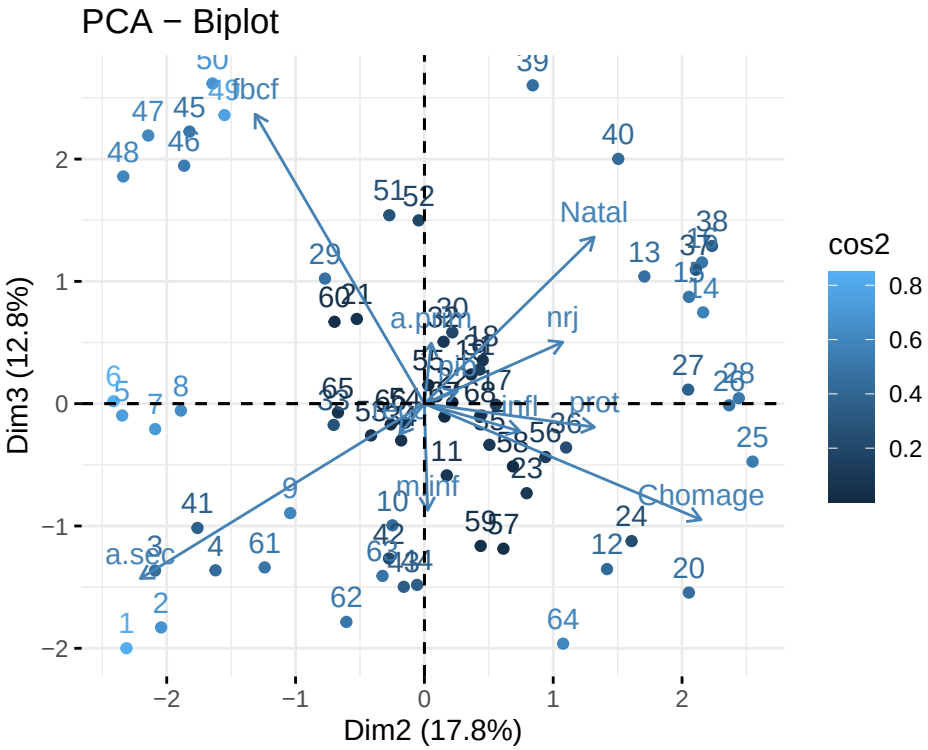


La variable formation brute de capital fixe (fcbf) était mal représentée sur l'axe (1,2) mais on s'aperçoit qu'elle est mieux représentée sur l'axe 3, donc l'axe 3 correspond à la formation brute de capital fixe.



On voit sur le plan constitué des deux premiers axes que le pays correspondant aux points allant de 57 à 60, en conséquent la Pologne est très bien représentée car il est éloigné du centre. On peut voir que son PIB est faible mais aussi que son taux de mortalité infantile ainsi que son pourcentage d'actif dans le secteur primaire

est très élevé. De plus on voit que le pays (5 à 8) donc l'Australie est bien représenté sur l'axe 2 tout comme (1 à 4) AL qui ont donc pourcentage d'actifs dans le secteur secondaire élevé contrairement aux pays (13 à 16) Canada et (25 à 28) EU qui ont un pourcentage d'actifs dans le secteur secondaire faible. D'autres pays sont quant à eux très mal représenté comme la Finlande (29 à 32) et la Russie (61 à 64) car les points se situent tous proches du centre.

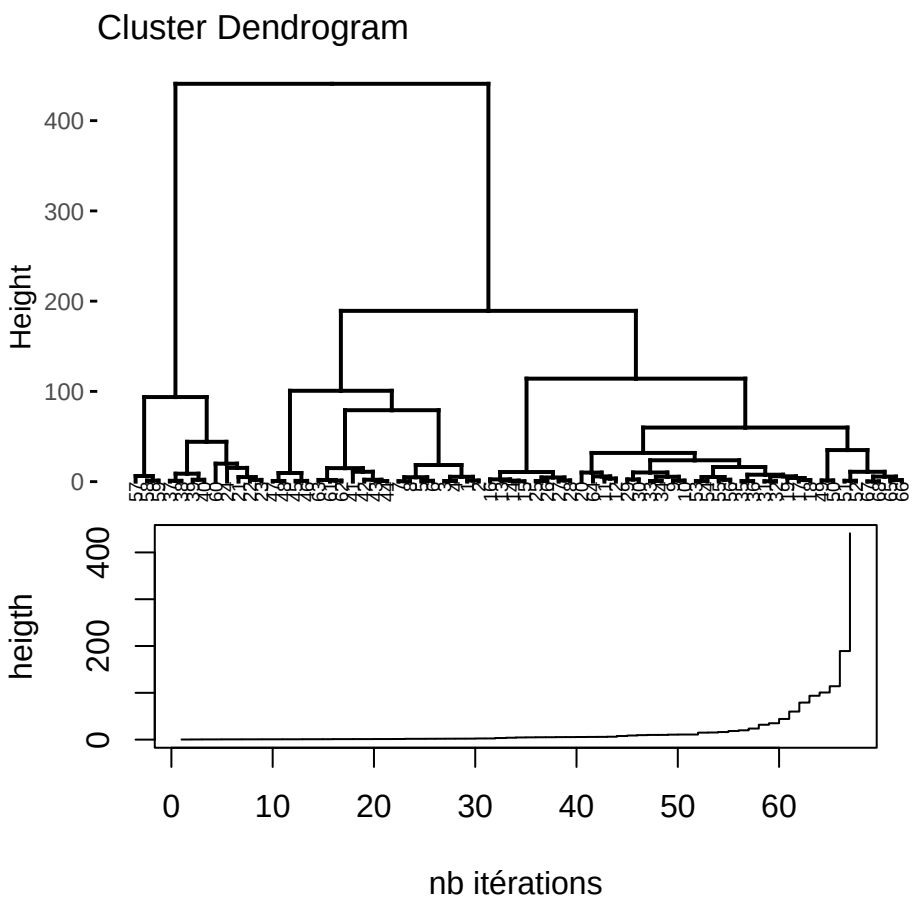


Sur ce graphe on peut voir que les pays (45 à 48) Japon et (49 à 52) Norvège ont une forte formation brute de capital fixe (axe 3) contrairement à un pays comme la Russie (61 à 64). On a que également que le pays (65 à 68) Suède est très mal représenté sur ce plan. On ne peut pas l'interpréter.

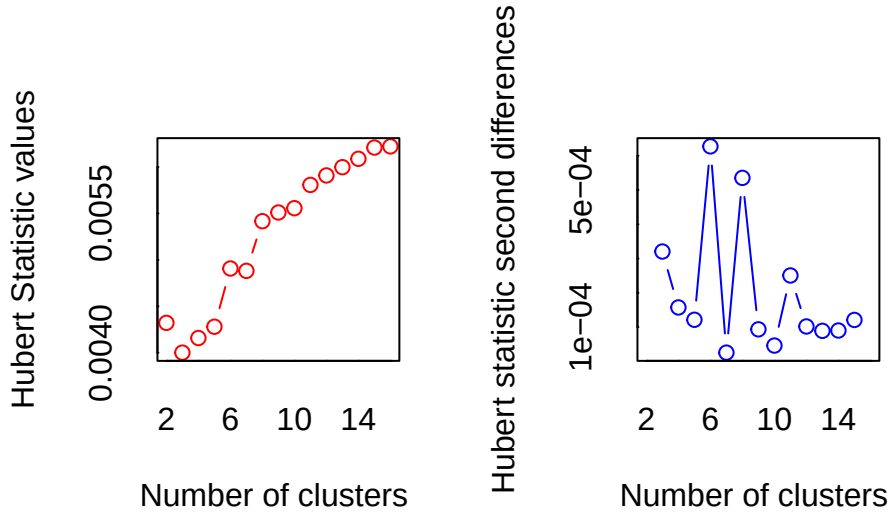
Dans ce contexte, une ACP simple n'est pas adaptée en raison des observations répétées pour chaque pays sur différentes années, introduisant ainsi une dépendance entre les observations.

Quatrième Partie : Classification non supervisée

```
## Warning: The `<scale>` argument of `guides()` cannot be `FALSE`. Use "none" instead as
## of ggplot2 3.3.4.
## i The deprecated feature was likely used in the factoextra package.
##   Please report the issue at <https://github.com/kassambara/factoextra/issues>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

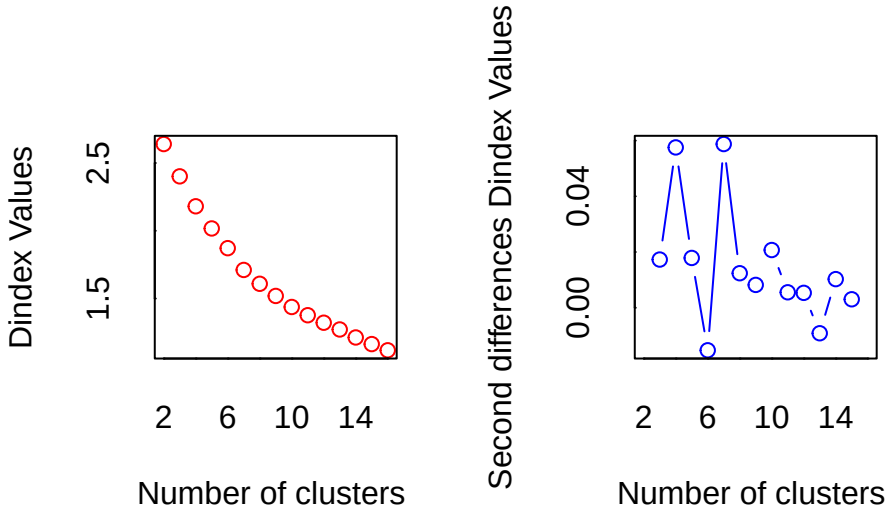


A l'étape initiale, l'inertie intra-groupes est nulle et va augmenter à chaque regroupement jsuqu'à atteindre l'inertie totale. L'augmentation de l'inertie intra est croissante en fonction du nombre d'itérations. Un grand saut dans ce graphe signfie qu'il y a une augmentation forte de l'inertie intra. C'est le cas ici quand on passe de 1 à 2 groupes de 2 à 3 et de 3 à 4 groupes. Ce graphe peut-être une aide au choix du nombre de groupes: on choisit un nombre de groupes associé au nombre de grand saut; ici, 3 ou 4 groupes.

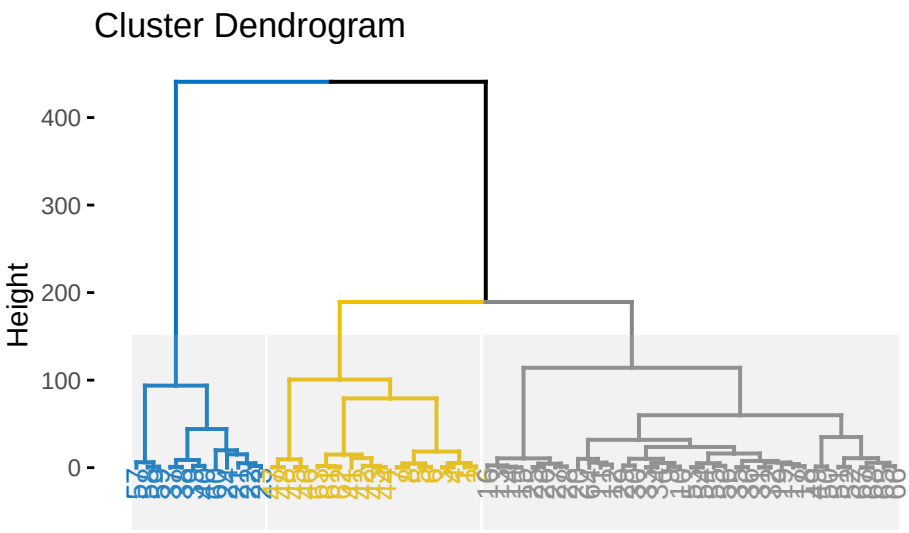



```
## *** : The Hubert index is a graphical method of determining the number of clusters.
##                In the plot of Hubert index, we seek a significant knee that corresponds to a
##                significant increase of the value of the measure i.e the significant peak in Hubert
```

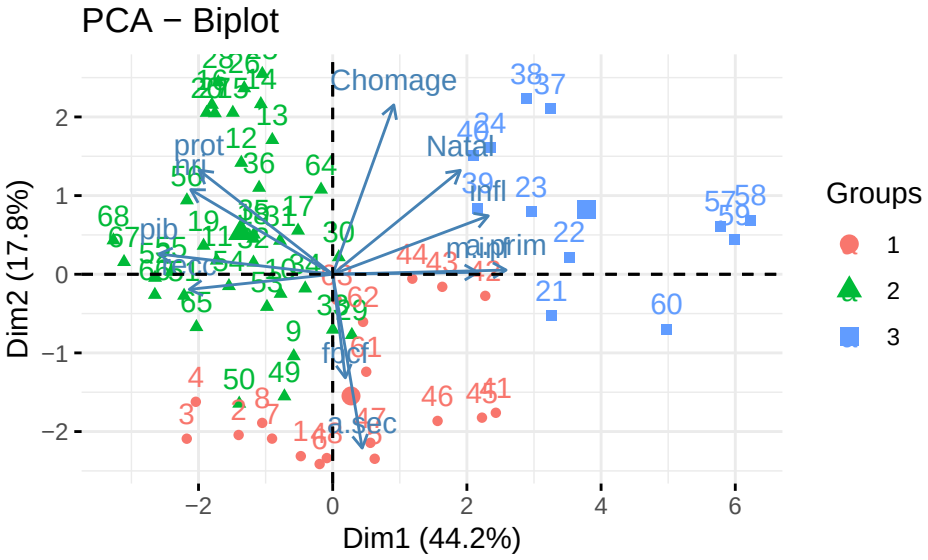
```
##                 index second differences plot.
##
```




```
## *** : The D index is a graphical method of determining the number of clusters.
##                 In the plot of D index, we seek a significant knee (the significant peak in Dindex
##                 second differences plot) that corresponds to a significant increase of the value of
##                 the measure.
##
## *******************************************************************
## * Among all indices:
## * 7 proposed 2 as the best number of clusters
## * 6 proposed 3 as the best number of clusters
## * 2 proposed 4 as the best number of clusters
## * 1 proposed 7 as the best number of clusters
## * 1 proposed 9 as the best number of clusters
## * 1 proposed 11 as the best number of clusters
## * 2 proposed 12 as the best number of clusters
## * 4 proposed 16 as the best number of clusters
##
##                    ***** Conclusion *****
##
## * According to the majority rule, the best number of clusters is  2
##
##
## *******************************************************************
```

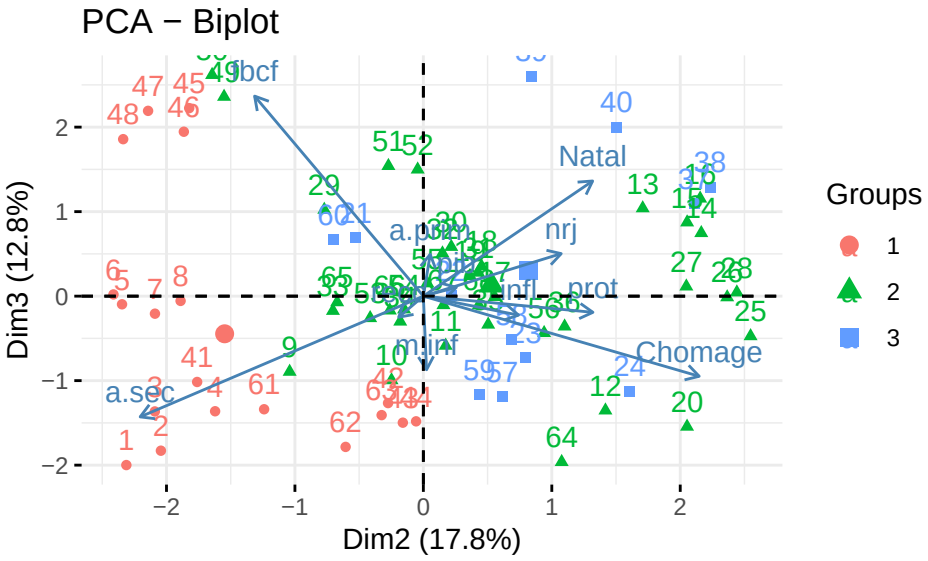


Les critères sélectionnent en majorité 3 groupes. On en choisit 3 ici. On peut représenter le dendrogramme avec ces 3 groupes. (graphe ci-dessus)



On observe que sur le graphe ci-dessus, les observations ont été regroupées en trois clusters distincts par la classification ascendante hiérarchique (CAH). Dans le biplot de l'analyse en composantes principales (ACP) ci-dessus, nous pouvons clairement distinguer une séparation entre ces groupes le long des axes 1 et 2. Les observations du groupe 2 sont principalement situées du côté négatif de l'axe 1, tandis que celles du groupe 3 sont majoritairement du côté positif de cet axe (Pays potentiellement plus pauvre et moins développé) . En revanche, les observations du groupe 1 se retrouvent principalement du côté négatif de l'axe 2 donc le groupe 1 à un pourcentage d'actifs dans le secteur secondaire élevé. Ces observations suggèrent des différences significatives dans les caractéristiques mesurées entre les groupes identifiés.

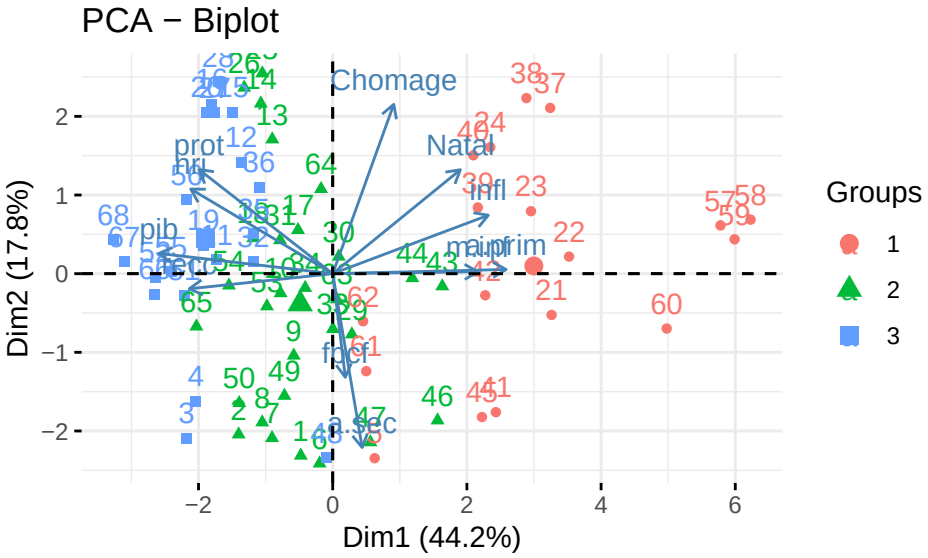
On voit que dans le groupe 3 il y a l'Espagne, le Portugal, Irlande cela correspond bien à nos suppositions.



Sur le graphe ci dessus, en dehors des observations du groupe 1 on ne peut pas classifier les observations en fonction de l'un des deux axes.

Group.1	Natal	Chomage	a.prim	a.sec	pib	fbcf	infl	recc	m.inf	prot	nrj
1	123.7	38.1	97.2	400.5	6892.5	236.7	99.3	374.3	143.4	52.9	336.8
2	133.9	53.8	70.6	329.2	9337.2	221.7	94.6	450.9	107.2	66.8	550.4
3	185.2	79.3	235.5	341.5	3218.4	232.0	172.1	302.2	212.5	53.0	177.8

On observe bien que le groupe 3 correspond aux pays avec les variables a.prim, Natal, chômage, inflation et mortalité infantile élevées ainsi que PIB et nrj faible donc des pays plus pauvres.



On aperçoit qu'il y a des chnagements de groupe pour certains pays mais nous allons comparer leurs différences pour mieux le voir.

```
##                cluster.CAH
## cluster.kmeans  1  2  3
##              1  6  0 12
##              2 10 19  0
```

```
##             3  3 18  0
```

```
## [1] 0.6637401
```

On sait que plus le rand index est proche de 1, plus les classifications sont similaires. Ici, on a donc que les classifications sont différentes mais ont quand même des sous-groupes communs.

Pour conclure, notre analyse des données socio-économiques des pays membres de l'OCDE sur la période 1975-1981 a mis en lumière plusieurs tendances. Nous avons observé une corrélation entre le PIB par habitant et le taux de chômage, soulignant l'impact du développement économique sur l'emploi. L'analyse en composantes principales a révélé des axes principaux reflétant les différences de développement entre les pays, tandis que la classification non supervisée a permis d'identifier des groupes de pays partageant des caractéristiques socio-économiques similaires. En conclusion, cette analyse fournit des informations pour comprendre les disparités et les dynamiques de développement au sein de l'OCDE, essentiels pour orienter les politiques publiques et les initiatives de croissance économique.