Apprentissage non-supervisé

Haeji Yun

Dans ce projet, nous appliquerons des méthodes d'apprentissage non-supervisé pour analyser les données associées à l'ouvrage *Une histoire du conflit* par *Julia Cagé* et *Thomas Piketty*.

Nous nous concentrerons sur les données de votes, de revenus, d'âges, de genre, d'éducation et de catégories socio-professionnelles des communes françaises. Les méthodes utilisées incluront la réduction de dimension pour visualiser les données, le clustering pour comprendre les caractéristiques des communes similaires, et la détection de ruptures pour analyser l'évolution des revenus au fil des années.

1. Dataset

Pour notre étude, nous avons créé le jeu de données pour chaque année où l'élection présidentielle a eu lieu. Les années concernées sont 1981, 1988, 1995, 2002, 2007, 2012, 2017, et 2022.

Pour cela, nous avons regroupé les variables quantitatives suivantes pour chaque commune :

- Le nombre et le pourcentage de vote pour chaque partie politique
- Le revenu moyen
- L'âge moyenne d'homme et de femme
- Le pourcentage d'obtention de bac
- Le pourcentage de propriétaire
- Le pourcetage de chaque catégorie socio-professionnelle
- Le nombre de population

Nous avons également créé des variables qualitatives suivantes à partir des variables quantitatives précédantes :

- Le parti politique majoritaire voté
- L'appartenance à une quartile du revenu moyen
- L'appartenance à une quartile d'obtention du baccalauréat
- L'appartenance à une quartile de propriétaires

- La catégorie socio-professionnelle majoritaire
- L'appartenance à une quartile de la population

Ces variables nous permettront de réaliser des analyses et de mieux comprendre les dynamiques socio-économiques et politiques au sein des communes françaises.

dep	10	10	10	10	10
codecommune	10002	10003	10004	10005	10006
inscrits	198.0	2640.0	159.0	205.0	1859.0
votants	154.0	1978.0	118.0	154.0	1309.0
exprimes	152.999996	1940.0	115.0	146.0	1278.0
voteG	35.599998	387.39999	12.8	21.4	221.60001
voteCG	4.6	60.400002	1.8	5.4	43.599998
voteC	41.599998	452.39999	19.799999	32.400002	315.60001
voteCD	4.6	101.4	11.8	7.4	74.599998
voteD	66.599998	938.40002	68.800003	79.400002	622.59998
pvoteG	0.23268	0.199691	0.111304	0.146575	0.173396
pvoteCG	0.030065	0.031134	0.015652	0.036986	0.034116
pvoteC	0.271895	0.233196	0.172174	0.221918	0.246948
pvoteCD	0.030065	0.052268	0.102609	0.050685	0.058372
pvoteD	0.435294	0.483711	0.598261	0.543836	0.487167
vote_majoritaire	pvoteD	pvoteD	pvoteD	pvoteD	pvoteD
votants/inscrits	0.777778	0.749242	0.742138	0.75122	0.704142
participation	Q2	Q1	Q1	Q1	Q1
revmoyadu2022	27245.238	14312.726	22348.033	18185.273	20139.592
revmoyfoy2022	31878.781	26175.957	30549.148	28732.73	24741.219
revenu	Q4	Q1	Q3	Q2	Q2
ageh2022	41.547371	37.624462	45.375835	37.782257	43.432163
agef2022	45.847401	47.47234	41.32008	39.884613	48.912537
pbac2022	0.179487	0.364837	0.519737	0.152284	0.256749
diplome	Q1	Q2	Q3	Q1	Q1
ppropri2022	0.72549	0.630709	0.868687	0.77686	0.516833
proprietaire	Q1	Q1	Q4	Q2	Q1
pagri2022	0.0	0.012248	0.509804	0.0	0.065
pindp2022	0.04878	0.092219	0.0	0.0	0.068333
pcadr2022	0.00813	0.154179	0.0	0.0	0.101667
ppint2022	0.317073	0.255043	0.058824	0.581395	0.125
pempl2022	0.455285	0.245677	0.392157	0.077519	0.333333
pouvr2022	0.170732	0.240634	0.039216	0.341085	0.306667
csp	pempl2022	ppint2022	pagri2022	ppint2022	pempl2022
pop2022	220.0	3514.0	178.0	242.0	2763.0
population	Q2	Q4	Q1	Q2	Q4

2. Réduction de dimension

Nous allons commencer par effectuer la réduction de dimensions pour visualiser et comprendre les structures sous-jacentes dans les données. Nous avons choisi l'année 2022 et nous allons utiliser les techniques de PCA, t-SNE et UMAP.

Pour le PCA et le UMAP, la normalisation des données a précédé l'application de la réduction de dimensions car ces techniques sont sensibles aux échelles des différentes variables.

2.1. Vote & Revenu

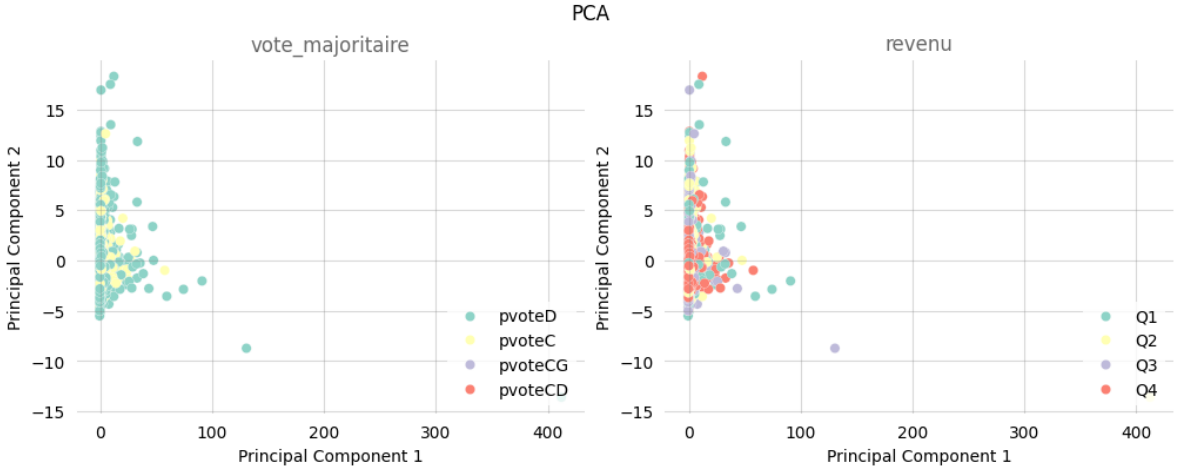
Dans un premier temps, nous allons nous concentrer uniquement sur les données de votes et de revenus. L'objectif est d'identifier des motifs ou des clusters potentiels qui pourraient révéler des liens entre les comportements électoraux et les niveaux de revenus des différentes communes.

dep	10	10	10	10	10
codecommune	10002	10003	10004	10005	10006
inscrits	198.0	2640.0	159.0	205.0	1859.0
votants	154.0	1978.0	118.0	154.0	1309.0
exprimes	152.999996	1940.0	115.0	146.0	1278.0
voteG	35.599998	387.39999	12.8	21.4	221.60001
voteCG	4.6	60.400002	1.8	5.4	43.599998
voteC	41.599998	452.39999	19.799999	32.400002	315.60001
voteCD	4.6	101.4	11.8	7.4	74.599998
voteD	66.599998	938.40002	68.800003	79.400002	622.59998
pvoteG	0.23268	0.199691	0.111304	0.146575	0.173396
pvoteCG	0.030065	0.031134	0.015652	0.036986	0.034116
pvoteC	0.271895	0.233196	0.172174	0.221918	0.246948
pvoteCD	0.030065	0.052268	0.102609	0.050685	0.058372
pvoteD	0.435294	0.483711	0.598261	0.543836	0.487167
vote_majoritaire	pvoteD	pvoteD	pvoteD	pvoteD	pvoteD
votants/inscrits	0.777778	0.749242	0.742138	0.75122	0.704142
participation	Q2	Q1	Q1	Q1	Q1
revmoyadu2022	27245.238	14312.726	22348.033	18185.273	20139.592
revmoyfoy2022	31878.781	26175.957	30549.148	28732.73	24741.219
revenu	Q4	Q1	Q3	Q2	Q2

PCA

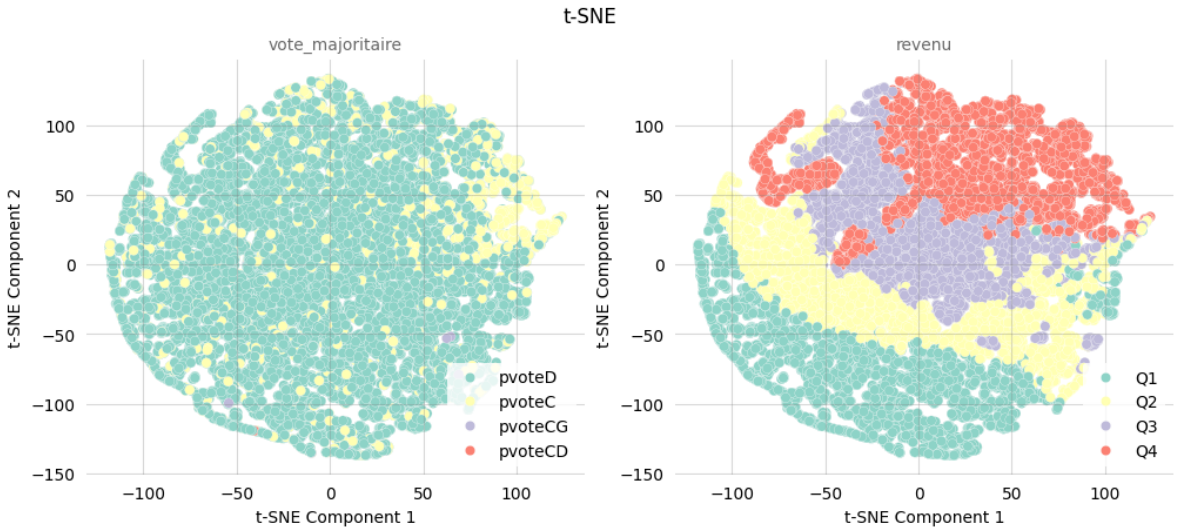
Les graphqiues ci-dessous montrent les résultats de PCA. Nous observons que la majorité des données se concentrent autour de l'origine sur la première composante principale avec une

dispersion plus importante le long de la deuxième composante principale. Les variables de vote et de revenu ne présentent pas de séparation claire dans l'espace PCA. Les deux premières axes principales ne distinguent pas efficacement le parti politique majoritaire voté et le niveau de revenu des communes.



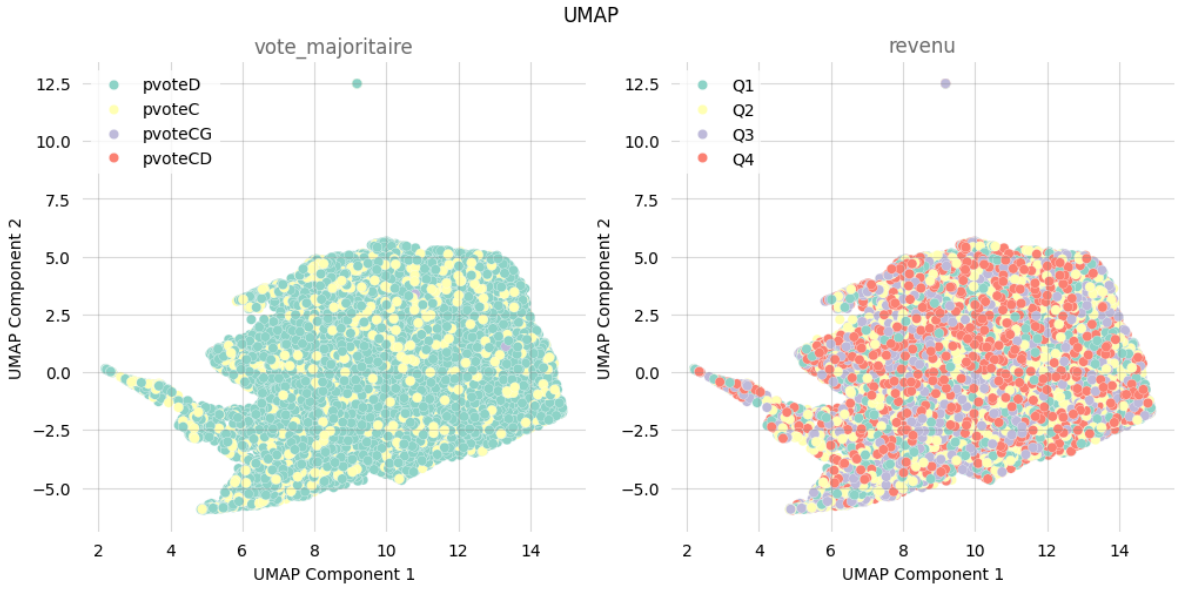
t-SNE

Quant à t-SNE, il montre une meilleure séparation des données par le niveau du revenu que par le parti politique majoritaire voté. Les communes à revenus faibles et élevés forment des clusters distincts, tandis que les votes restent dispersés sans regroupement clair. Les variables de revenu ont une influence plus marquée sur la structuration des données par rapport aux partis politiques votés.



UMAP

Le UMAP présente une dispersion uniforme des données sans séparation claire. Les points sont largement superposés, indiquant que UMAP n'a pas réussi à capturer des clusters distincts dans ces données.



Les données choisies ne semblent pas être linéairement réductibles, comme l'indique la performance limitée de la PCA. Bien que le UMAP n'ait pas réussi à distinguer de manière significative les clusters pour les votes et les revenus, le t-SNE a montré que le technique non linéaire qui préserve les relations de proximité entre les points, est plus efficace pour révéler les structures dans les données.

2.2. Toutes les variables

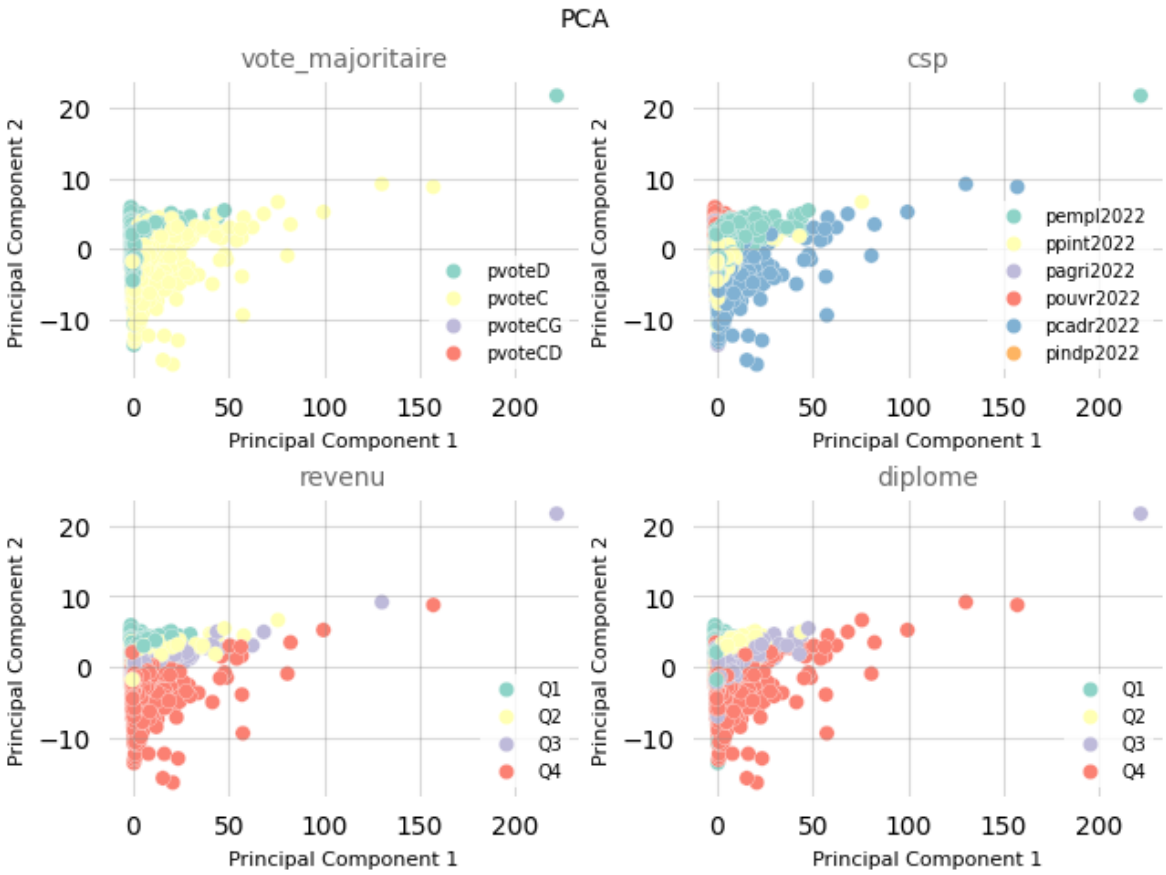
Nous élargirons notre analyse en incluant d'autres variables socio-économiques telles que l'âge moyen des habitants, le pourcentage de bacheliers, le pourcentage de propriétaires, et les catégories socio-professionnelles. Cela nous permettra d'obtenir une vision plus complète et plus nuancée des facteurs influençant les résultats électoraux et les conditions économiques des communes françaises.

PCA

En appliquant le PCA sur toutes les variables, nous observons des séparations distinctes pour les différentes variables, principalement le long de la composante principale 2.

Vote
La distinction est claire entre les votes de droite et les votes centristes.

CSP
La catégorie socio-professionnelle est visiblement bien distinguée tout au long de l'axe principal 2. Les catégories socio-professionnelles ouvrières montrent une tendance à voter pour le parti droit et les catégories cadres ont une tendance à voter pour le parti centriste. Les autres catégories socio-professionnelles montrent une distribution mélangée.



Revenu
Le niveau de revenu est négativement lié à la composante principale 2. Plus le revenu est bas, plus il se situe en haut de l'axe 2. Les communes à bas revenu (Q1 et Q2) montrent une certaine tendance à voter pour le parti droit. Les communes à revenu plus élevé (Q3 et Q4) sont plus favorables au parti central.

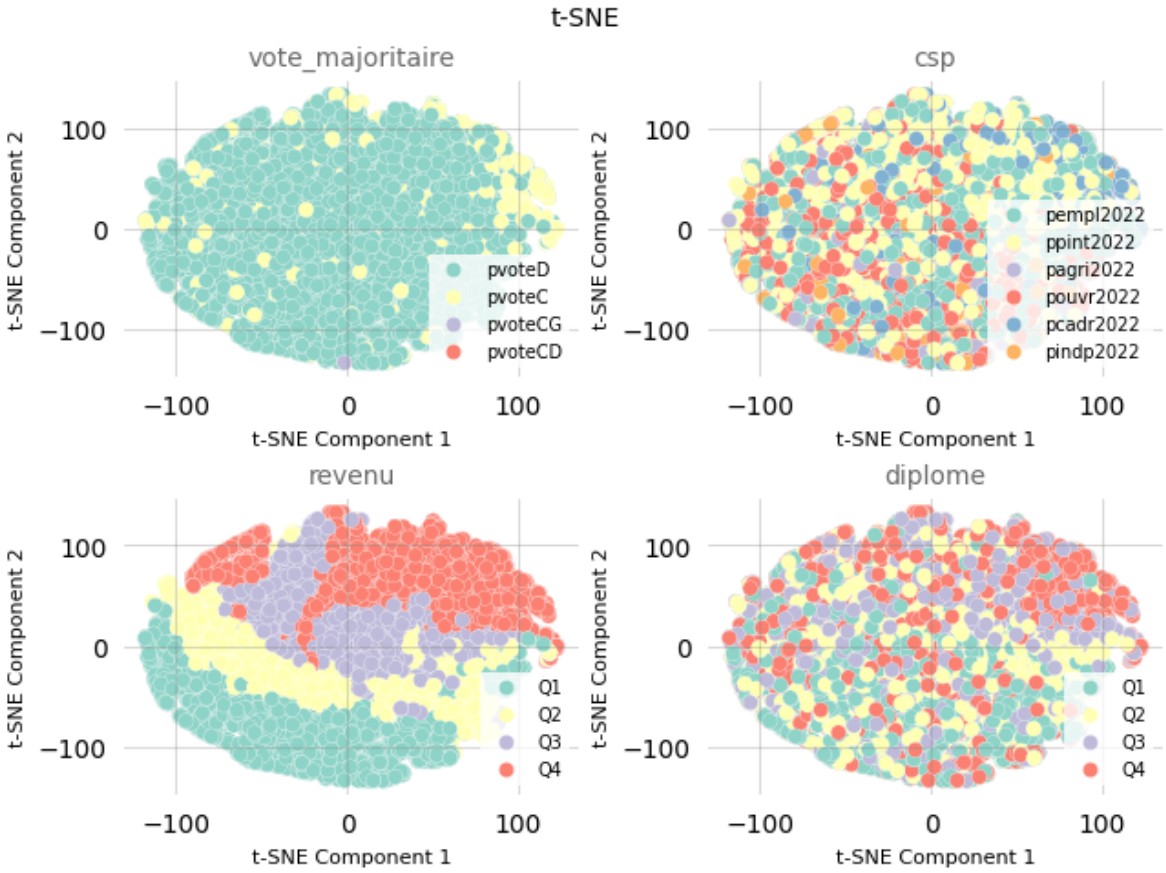
Education
Le pourcentage d'obtention de bac suit une tendance similaire au revenu, étant également négativement lié à la composante principale 2. Plus le pourcentage est bas, plus il se situe en

haut de l'axe 2. Comme le revenu, les communes avec un bas pourcentage d'obtention de bac (Q1 et Q2) montrent une certaine tendnace à voter pour le parti droit. Les communes avec le pourcentage d'obtention plus élevé (Q3 et Q4) ont un caractère centriste. Il est possible qu'il y ait des interactions entre le niveau d'éducation, le revenu, la professsion et les comportements de vote.

Avec l'intégration de toutes les variables, les données résument mieux la variabilité de manière linéaire. Les résultats révèlent également des relations entre les variables socio-économiques et politiques. Ces résultats suggèrent que les communes à bas revenu, avec un faible niveau d'obtention de bac et majoritairement ouvrières, sont principalement en faveur des partis de droite. D'un autre côté, les communes à revenu plus élevé, majoritairement composées de cadres, ont tendance à voter pour les partis centristes.

t-SNE

Quant à t-SNE, le niveau de revenu influence visiblement la réduction de dimension. Nous observons une distinction nette entre les différents niveaux de revenu.

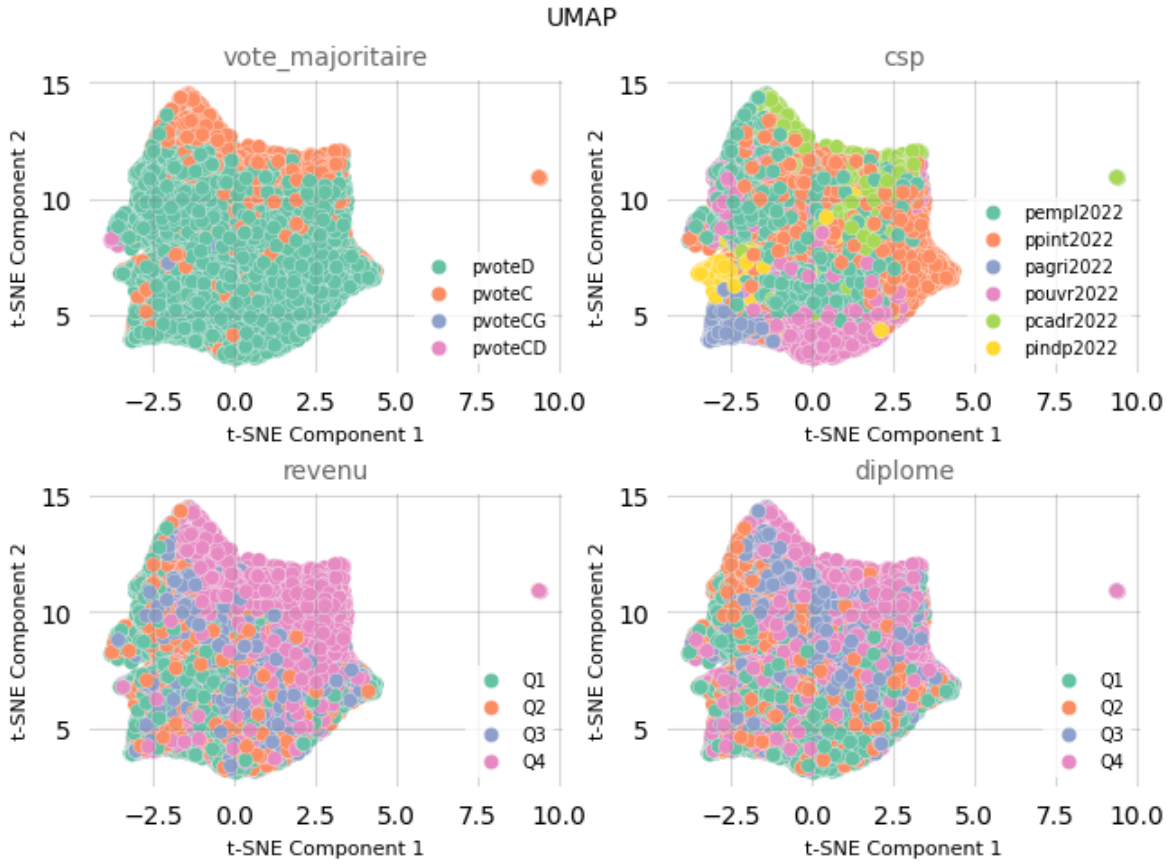


Le reste de variables est très dispersé avec quelques légères tendances : les ouvriers sont plus concentrés vers la gauche et le bas du graphique, tandis que les cadres se trouvent plus vers la droite et le haut. Les communes avec un pourcentage élevé d'obtention de bac ont tendance à se regrouper vers le haut du graphique.

Néanmoins, il est difficile de dégager une supposition claire sur le lien entre les variables. Les données sont moins bien représentées en basse dimension avec la méthode non linéaire en présence de toutes les variables.

UMAP

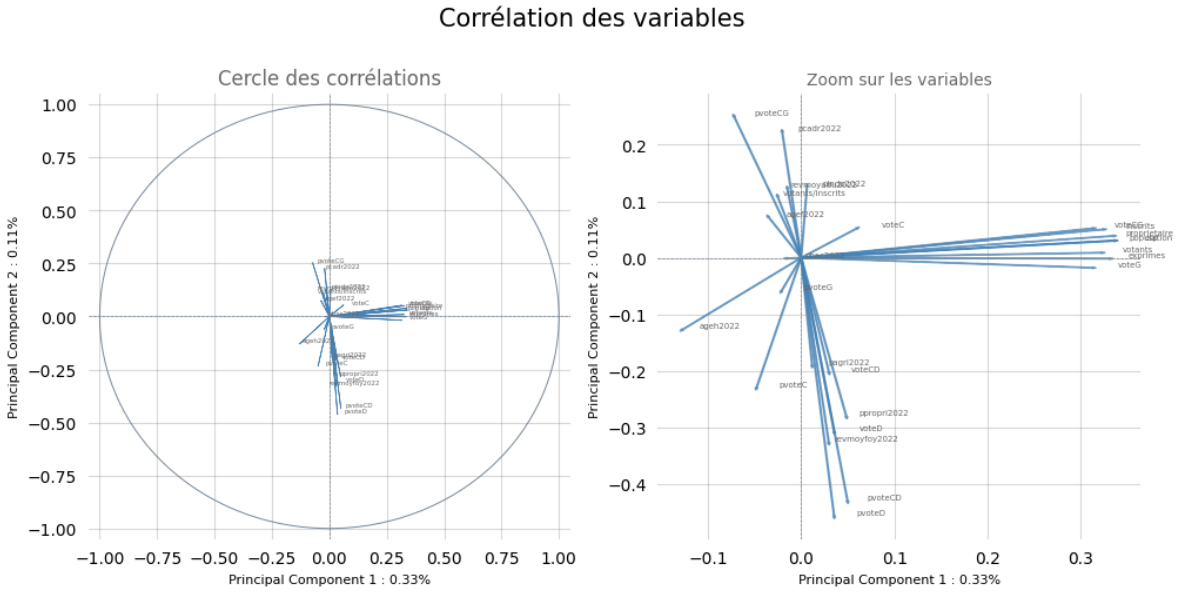
Avec le UMAP, malgré des tendances légères qui apparaissent, les communes sont largement superposées sans séparation claire par catégorie de chaque variable. Il est difficile de dégager une compréhension approfondie de nos données avec UMAP.



Nous allons garder les résultats de PCA pour les analyses car elle représente le mieux les données en basse dimension et est la méthode la plus informative parmi celles appliquées.

2.3. Cercle de corrélations

Nous allons visualiser le cercle des corrélations pour montrer les contributions des variables.



Nous observons que les deux premières composantes expliquent une bonne partie de la variabilité des données, atteignant 70%. Par contre, la majorité des variables se trouvent près du centre du cercle. Chaque variable a une faible contribution aux deux premières composantes.

Composante Principale 1
Les variables comme le nombre de votes, le nombre de votants et la population contribuent fortement à cette composante. Par conséquent, les communes qui se trouvent vers la droite du graphique seraient principalement des grandes communes en termes de population et de nombre de votants

Composante Principale 2
Cette composante oppose des contributions contrastées, avec une forte contribution négative du pourcentage de votes pour le parti de droite et une forte contribution positive du pourcentage de cadres. Les communes favorables aux partis de droite et celles majoritairement composées de cadres ont tendance à se retrouver dans des côtés opposés. Parmi les autres contributions positives sur l'axe 2, nous trouvons également le pourcentage de participation au vote, le revenu moyen et l'âge des femmes. Parmi les contributions négatives sur l'axe 2, nous avons le pourcentage d'agriculteurs, le pourcentage de propriétaires et l'âge des hommes.

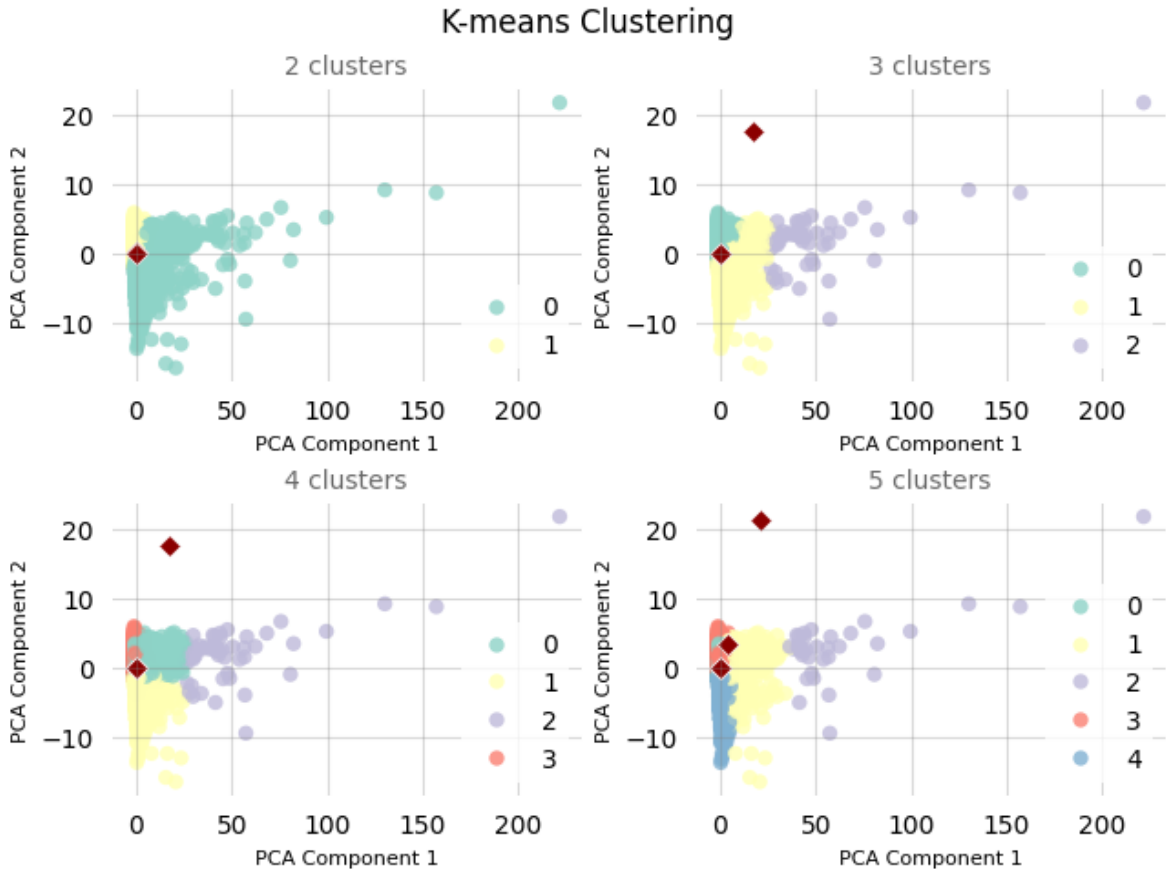
3. Clustering

Pour approfondir notre analyse, nous allons effectuer un clustering des données en utilisant l'algorithme K-means.

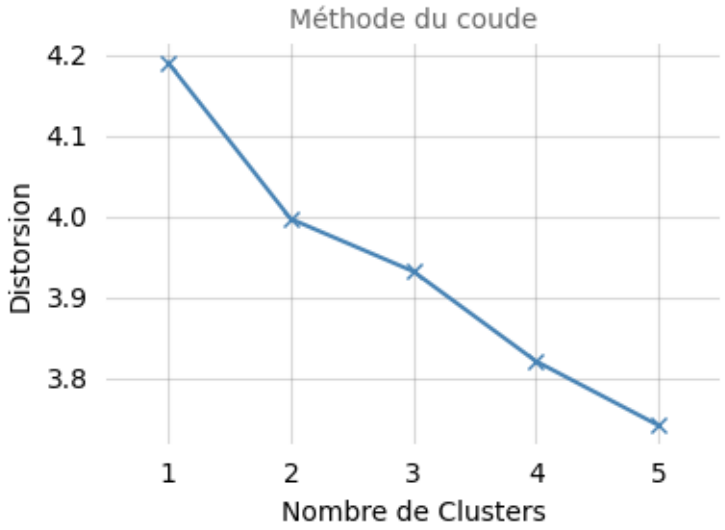
3.1. Clustering

Nous allons explorer différent nombre de cluster et les représenter sur le résultat de PCA. Les centres des clusters sont représentés par des points en losange rouge.

Nous observons déjà que la séparation en 2 clusters présents les caratéristiques similaires aux distinctions obtenues avec le PCA. Le cluster 0 correspond aux communes qui ont tendance à voter pour le parti droit avec des revenus bas et une forte proportion d'ouvriers. Le cluster 1 regroupe les communes avec des caractéristiques opposées, telles que des revenus plus élevés et une proportion plus importante de cadres.

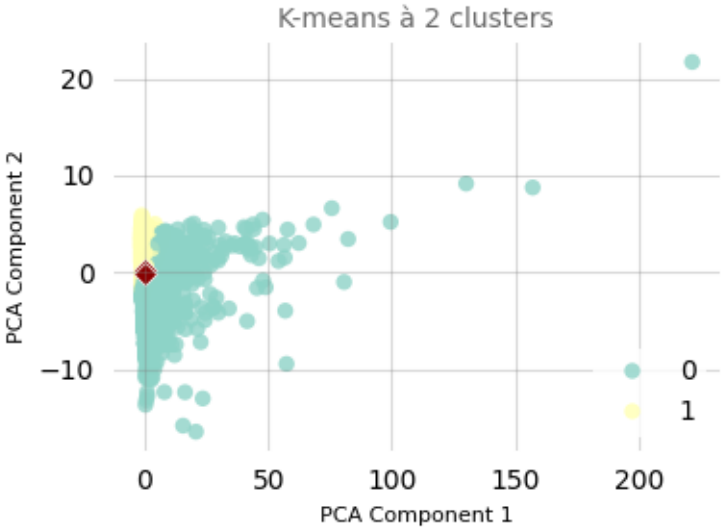


En utilisant la méthode du coude, nous pouvons déterminer que le nombre de cluster optimal est de 2 qui confirme notre observation initiale et valide l'utilisation de 2 clusters.



Cependant, le fait que les centres de clusters soient souvent superposés ou en dehors des données représentées indique que les deux premières composantes principales ne captureraient pas toute la variance importante des données et les clusters pourraient être influencés par des dimensions qui ne sont pas bien représentées par ces deux composantes.

Nous allons poursuivre notre analyse avec 2 clusters.



Un grand cluster 0 comprend la majorité des communes, réparties sur une large étendue des

composantes principales, et un petit cluster 1 comprend un nombre beaucoup plus réduit de communes, principalement regroupées autour de l'origine.

Le centre du cluster 1 est proche de l'origine et le centre du cluster 0 est légèrement décalé vers la droite. Ces centres sont partiellement superposés et se trouvent en dehors des zones de forte densité de données, ce qui indique que les deux premières composantes principales ne capturent pas toutes les informations nécessaires pour une séparation claire par K-means.

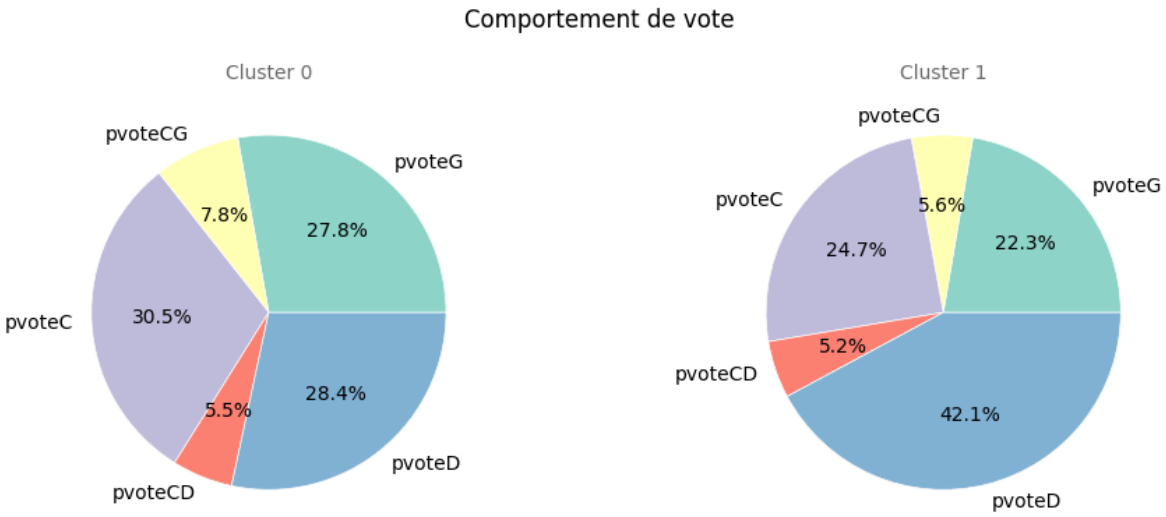
Le clustering K-means avec 2 clusters montre une séparation qui correspond aux distinctions observées avec le PCA. Le cluster 1 correspond aux communes favorables au parti de droite, avec une majorité d'ouvriers, un revenu bas et un faible pourcentage d'obtention de bac. Le cluster 0, quant à lui, correspond aux communes qui n'ont pas ces caractéristiques socio-économiques, ayant une préférence pour le parti central.

Bien que les deux premières composantes principales ne capturent pas entièrement la séparation effectuée par K-means, nous pouvons dégager une idée sur les caractéristiques des clusters: un lien entre la catégorie socio-professionnelle, le niveau de revenu et la préférence pour un parti politique.

3.2. Analyse de clusters

Nous pouvons étudier les caractéristiques des deux clusters plus en détail.

Vote
Le graphique ci-dessous montre la répartition des votes pour différents partis politiques dans les deux clusters.

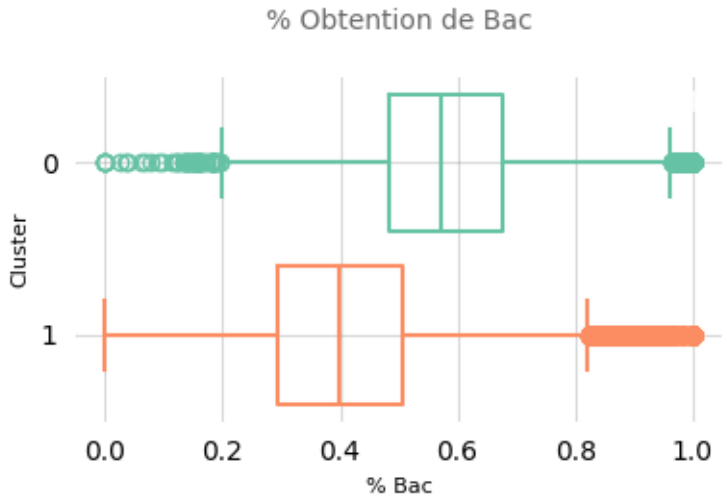


Dans le cluster 0, la majorité des votes se dirige vers le parti central avec 30%, suivi par le parti de droite et le parti de gauche, tous les deux autour de 28%. Les votes sont équitablement répartis entre les partis central, droit et gauche, ne présentant pas de caractère extrême. Le cluster 1 a une préférence visible pour le parti de droite, qui obtient 42% des votes, suivi par le parti central avec 25% et le parti de gauche avec 22%. Ce cluster montre une préférence graduelle de droite à gauche, avec un fort caractère droitiste.

Le comportement différent sur les dynamiques politiques est clairement visible dans les deux clusters.

Education
Quant au pourcentage d'obtention du baccalauréat, le cluster 0 regroupe les communes avec un pourcentage d'obtention du bac autour de 50% à 70%. Les communes ayant un taux d'obtention inférieur à 20% sont considérées comme outliers. Dans le cluster 1, le pourcentage d'obtention du bac est autour de 40%. Néanmoins, le cluster présente une hétérogénéité plus prononcée avec un nombre important de communes ayant un pourcentage d'obtention du bac supérieur à la moyenne du cluster.

Nous observons une nette différence de niveau d'éducation entre les deux clusters. Le cluster 0 a un pourcentage d'obtention du bac clairement plus élevé que le cluster 1.



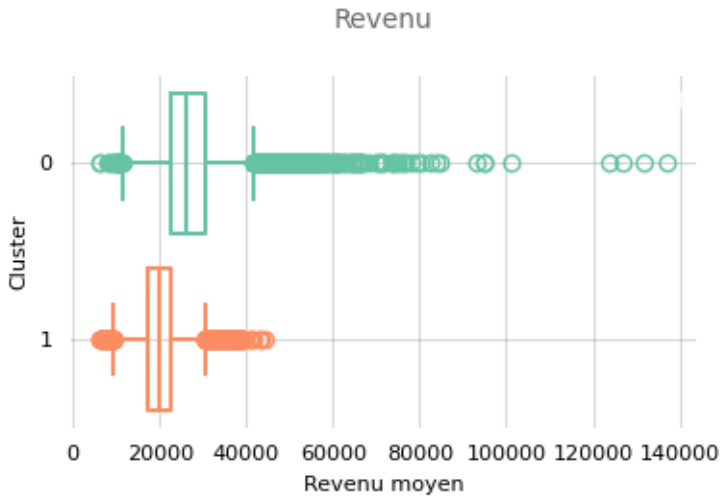
CSP
Nous pouvons également remarquer la différence dans la répartition des catégories socio-professionnelles.

Le cluster 0 est majoritairement composé de professions intermédiaires, d'employés et de cadres. Le cluster 1 est caractérisé par une forte présence d'employés, d'ouvriers et de professions inter-

médiaires. Bien que les professions intermédiaires et les employés soient fortement représentés dans les deux clusters, le cluster 1 se distingue par la présence prononcée des ouvriers.

Revenu
Au niveau du revenu, il y a une disparité significative entre les deux clusters.

Le cluster 0 présente un niveau de revenu plus élevé à l'entour de 25.000 euros avec une dispersion dans les tranches de revenus extrêmement hauts. Le cluster 1 comprend un niveau de revenu autour de 20.000 euros, marqué par des revenus moyens plus bas et une concentration plus étroite autour des valeurs médianes. Le custer présente une homogénéité économique avec moins de variation extrême.



L'analyse des deux clusters identifiés par le K-means révèle des différences socio-économiques et politiques significatives entre les communes. Le cluster 0 se distingue par une population plus diversifiée économiquement, avec des niveaux de revenu plus élevés, une répartition équilibrée des votes entre les partis politiques, et un pourcentage plus élevé d'obtention du baccalauréat. En revanche, le cluster 1 est caractérisé par des revenus moyens plus bas, une forte présence d'ouvriers et d'employés, et une préférence marquée pour le parti de droite

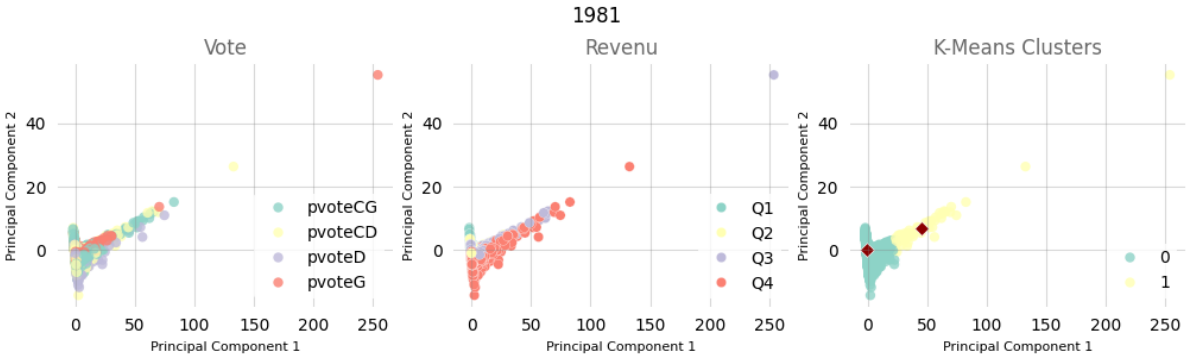
3.3. Evolution au cours des années

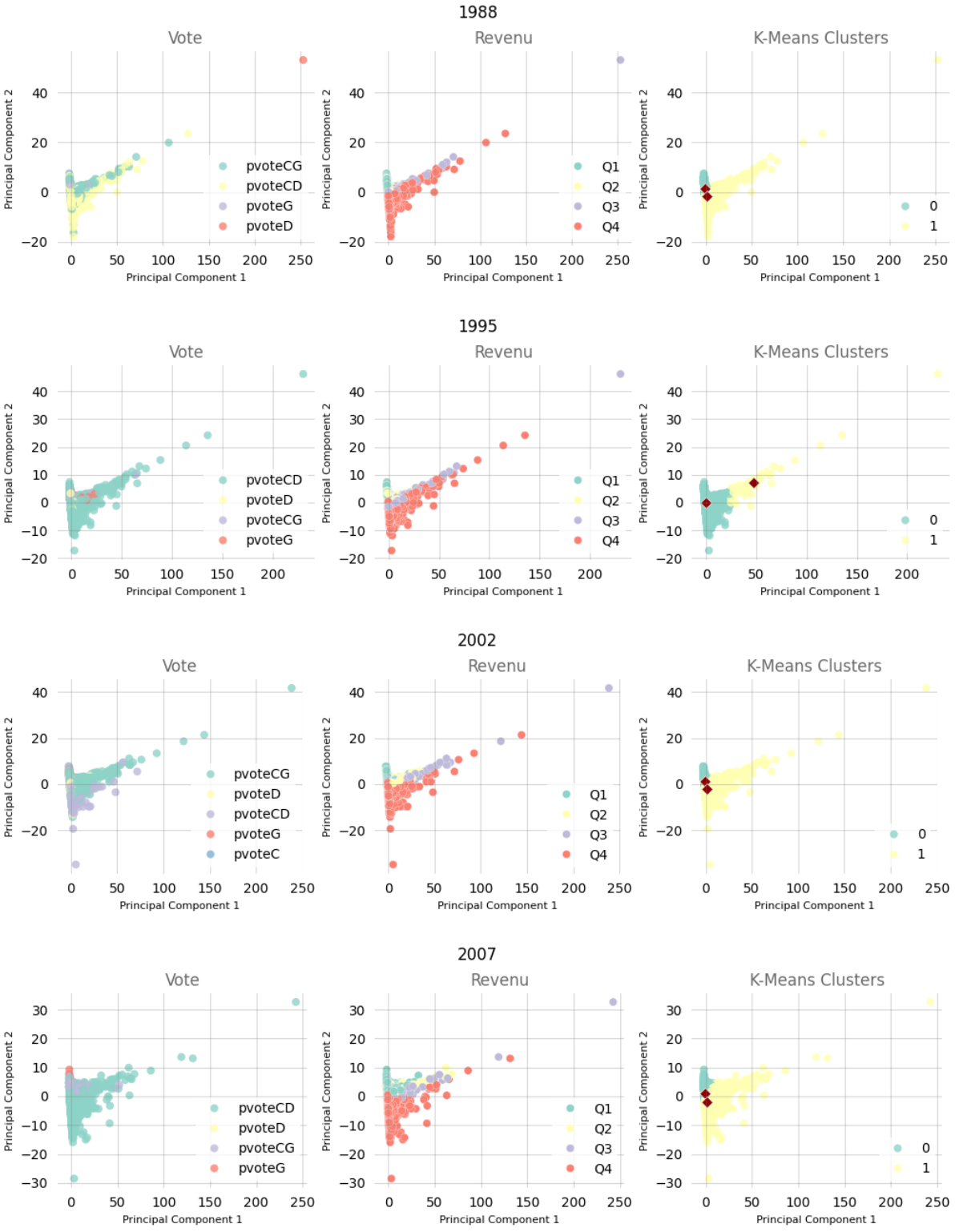
Nous allons appliquer le PCA et le K-means au cours des différentes années pour voir si on retrouve les mêmes caractéristiques et comment les clusters évoluent au fil du temps. Les graphiques ci-dessous montrent les clusters, le comprotement de vote et le niveau de revenu de chaque années de l'élection présidentielle.

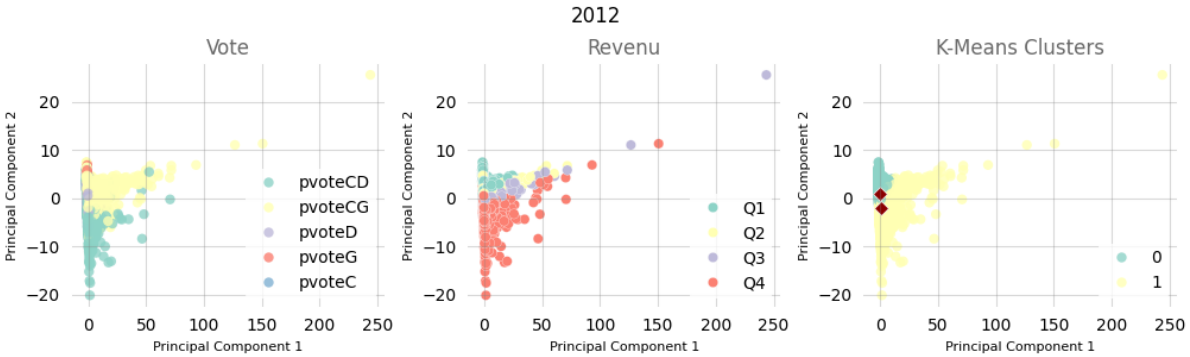
À l'exception des années 1981 et 1995, où il est difficile de dégager des caractéristiques claires pour les deux clusters, nous pouvons observer des traits similaires qu'en 2022 que nous venons d'analyser dans les autres années.

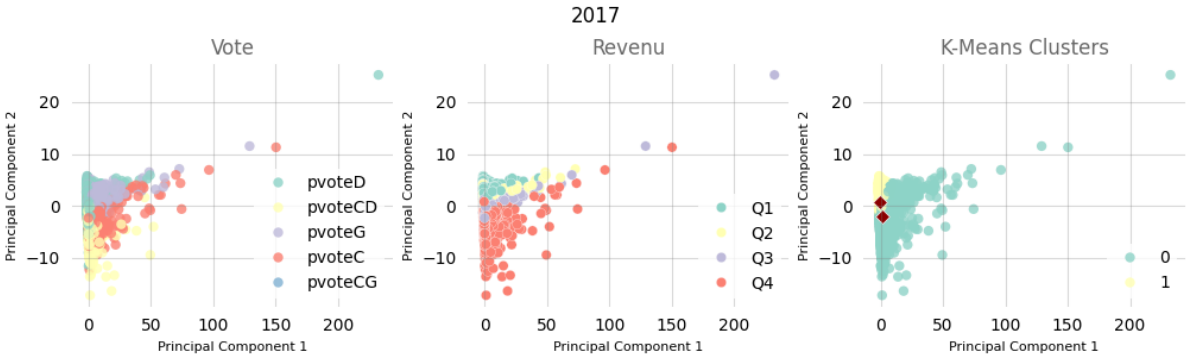
Dans les autres années, nous observons le petit des deux clusters regroupant les communes avec un niveau de revenu bas, avec une préférence envers les partis gauchistes et centre-gauchistes, avant de basculer vers le parti de droite à partir de 2017.

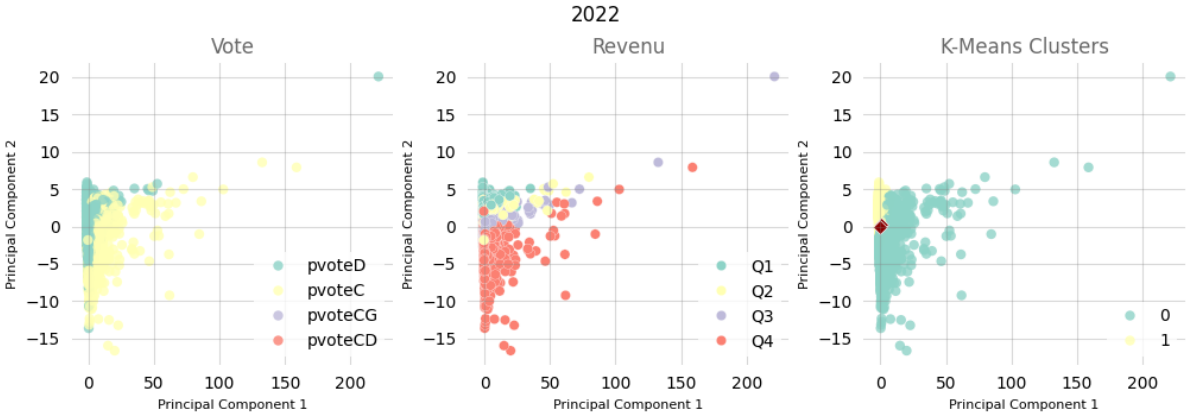
Globalement, les tendances observées en 2022 se maintiennent avec une dispersion continue des votes et des revenus. Les centres des clusters K-means montrent une stabilité relative avec une légère variation au fil des ans, indiquant des changements progressifs dans les caractéristiques des communes. Les analyses PCA et K-means révèlent des tendances claires et cohérentes dans la répartition des communes selon les votes et le revenus.




1988
Vote
Revenu
K-Means Clusters
Principal Component 2
Principal Component 1
pvoteCG
pvoteCD
pvoteG
pvoteD
Q1
Q2
Q3
Q4
0
1
1995
pvoteCD
pvoteD
pvoteCG
pvoteG
2002
pvoteC
2007

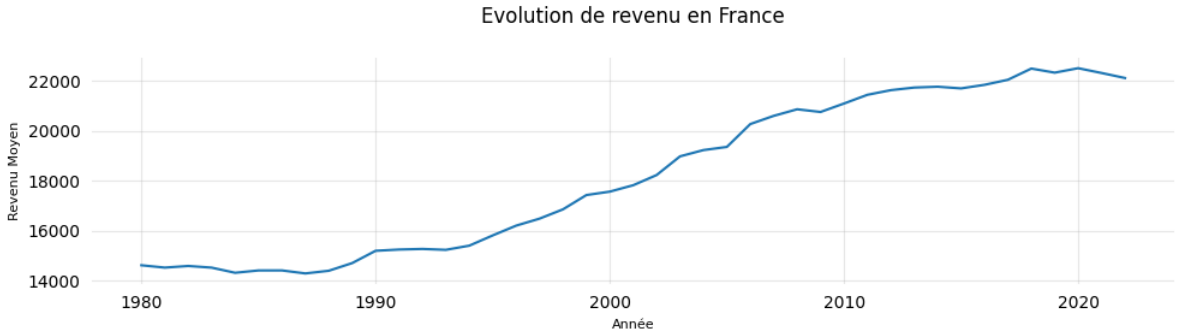

2012
Vote
Revenu
K-Means Clusters
Principal Component 2
Principal Component 1
pvoteCD
pvoteCG
pvoteD
pvoteG
pvoteC
Q1
Q2
Q3
Q4
0
1


2017
Vote
Revenu
K-Means Clusters
Principal Component 2
Principal Component 1
pvoteD
pvoteCD
pvoteG
pvoteC
pvoteCG
Q1
Q2
Q3
Q4
0
1


2022
Vote
Revenu
K-Means Clusters
Principal Component 2
Principal Component 1
pvoteD
pvoteC
pvoteCG
pvoteCD
Q1
Q2
Q3
Q4
0
1

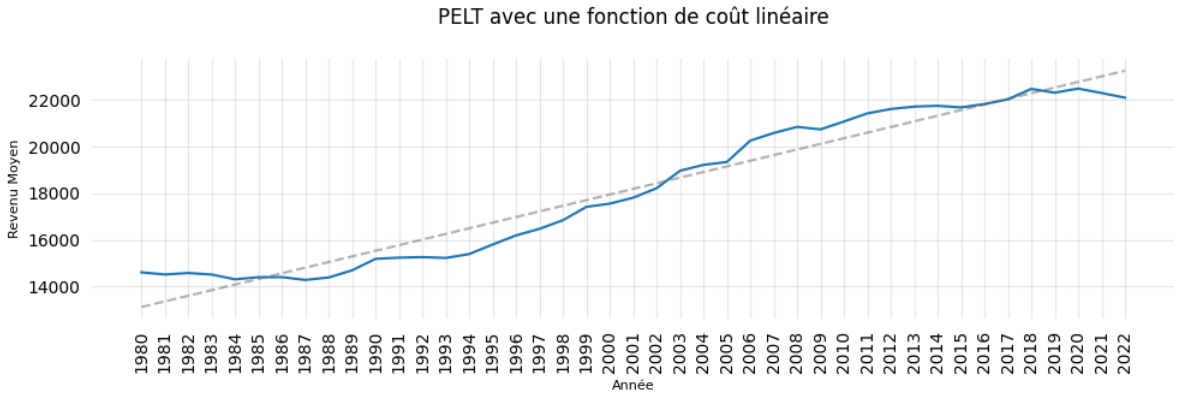
4. Détection de rupture

Maintenant nous allons nous concentrer sur le revenu. L'évolution du revenu moyen en France de 1980 à 2022 montre une tendance générale à la hausse avec des variations notables à certains moments. Pour mieux comprendre ces variations et identifier les périodes de changement significatif dans la dynamique des revenus, nous allons appliquer une méthode de détection de ruptures.



4.1. Détection de rutpure avec fonction de coût linéaire

Dans un premier temps, nous allons appliquer la méthode PELT avec une fonction de coût linéaire. Nous observons que le modèle ne détecte aucune rupture significative. Selon le modèle, l'évolution des revenus n'a pas connu de variations abruptes mais plutôt une progression linéaire représentée par la courbe grise sur le graphique. Cela indique une stabilité relative dans les facteurs économiques influençant le revenu moyen, sans événements perturbateurs majeurs ayant un impact direct sur la tendance générale des revenus.



4.1. Détection de rutpure avec fonction de coût non-linéaire

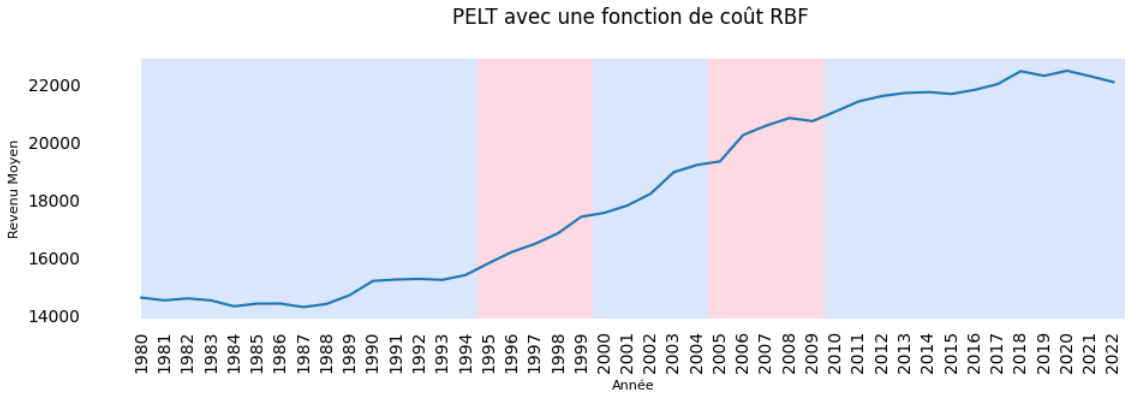
Maintenant, nous allons appliquer la méthode PELT avec une fonction de coût non-linéaire, RBF. Avec la fonction de coût non-linéaire, nous avons identifié quatre points de rupture. Les zones colorées indiquent les segments où des ruptures ont été détectées.

Un premier point de rupture est détectée en 1995, marquant le début d'une période de légère croissance par rapport aux années précédentes.

Un deuxième point de rupture est détectée en 2000, indiquant une accélération de la croissance des revenus.

À partir du troisième point de rupture en 2005, on observe une stabilité relative après l'augmentation rapide des revenus dans les années précédentes.

Un dernier point de rupture de 2009 correspond à la crise économique mondiale, après laquelle l'augmentation du revenu stagne.



L'application des méthodes linéaire et non-linéaire de détection de ruptures a mis en évidence des différences dans l'analyse des données de revenu moyen. La méthode linéaire suggère une stabilité continue, tandis que la méthode non-linéaire révèle des ruptures, indiquant des périodes de changement économique notable.

5. Conclusion

Cette étude a permis d'analyser les données socio-économiques et politiques des communes françaises en utilisant diverses techniques d'apprentissage non supervisé, incluant la réduction de dimension, le clustering et la détection de ruptures.

En appliquant les méthodes du ACP et du K-means, nous avons identifié deux principaux clusters de communes françaises aux caractéristiques socio-économiques et politiques distinctes.

Ces clusters ont été visualisés en deux dimensions, permettant une compréhension claire des différences entre ces groupes.

La détection de ruptures a permis d'identifier les moments de changements dans l'évolution des revenus, en considérant à la fois des perspectives linéaires et non-linéaires.