AFC et Classification

Haeji Yun

2023-05-01

Dans ce projet, nous étudions les résultats de l'élection présidentielle 2017. Pour cela, nous allons analyser le nombre de votes que les candidats ont obtenu à chaque département.

1. Analyse Préliminaire

Les données sont composées de 106 observations et 14 variables sans valeurs manquantes. Les observations sont les départements et les variables sont les différentes sorties que nous pouvons obtenir au vote : le nom des 11 candidats présidentiels, l'abstention, le blanc et le nul.

En effet, c'est le tableau de contingence formé par le département et le candidat (ou la sortie du vote). Le tableau donne les fréquences formées par les deux variables.

Le tableau ci-dessous montre l'aperçu de données.

```
##                          Abstentions Blancs Nuls LE PEN MELENCHON MACRON FILLON
## Ain                            81530   6342 2239  81455     51736  73692  69804
## Aisne                          80183   5047 2323 102787     48959  51693  46985
## Allier                         54357   4240 2556  43071     38324  45744  36499
## Alpes-de-Haute-Provence        24323   1806  736  24463     22448  19960  18442
## Alpes-Maritimes               161905   9029 2907 163140     87941 111943 161035
## Ardèche                        46619   3901 1667  45588     42880  42703  34182
##                          LASSALLE DUPONT-AIGNAN HAMON ASSELINEAU POUTOU ARTHAUD
## Ain                          3465         19788 16711       3612   3098    1842
## Aisne                        2265         14652 12231       2171   3156    2764
## Allier                       2988          9819 10639       1483   2328    1543
## Alpes-de-Haute-Provence      1721          4860  4983        932   1178     521
## Alpes-Maritimes              5262         25175 21067       6067   3622    1729
## Ardèche                      3581          9994 11844       1985   2623    1317
##                          CHEMINADE
## Ain                            595
## Aisne                          536
## Allier                         355
## Alpes-de-Haute-Provence        205
## Alpes-Maritimes                939
## Ardèche                        376
```

Dans le résumé ci-dessous de notre jeu de données, nous pouvons remarquer quelques informations intéressantes.

Nous constatons un nombre d'abstention assez important et grande variabilité entre les candidats.

Nous observons pour chaque candidat qu'il y a un grand écart entre le minimum et la première quartile. Cela suggère que les candidats ont au moins un département où ils ont reçu particulièrement peu de votes par rapport aux autres départements.

Nous observons également un grand écart entre la troisième quartile et le maximum pour tous les candidats. Les candidats ont au moins un département où ils ont obtenu particulièrement beaucoup de votes.

```
##   Abstentions         Blancs           Nuls           LE PEN
##  Min.   :  2238   Min.   :   35   Min.   :   26   Min.   :   380
##  1st Qu.: 39437   1st Qu.: 3170   1st Qu.: 1387   1st Qu.: 32646
##  Median : 73781   Median : 5090   Median : 2342   Median : 60658
##  Mean   : 93145   Mean   : 6154   Mean   : 2715   Mean   : 72107
##  3rd Qu.:119657   3rd Qu.: 8901   3rd Qu.: 3355   3rd Qu.: 91828
##  Max.   :417073   Max.   :24060   Max.   :13305   Max.   :382030
##    MELENCHON          MACRON          FILLON          LASSALLE
##  Min.   :   192   Min.   :   473   Min.   :   261   Min.   :   29
##  1st Qu.: 26328   1st Qu.: 30898   1st Qu.: 25965   1st Qu.: 1953
##  Median : 50390   Median : 58640   Median : 51045   Median : 3360
##  Mean   : 65779   Mean   : 79559   Mean   : 66676   Mean   : 4084
##  3rd Qu.: 81216   3rd Qu.:107690   3rd Qu.: 95068   3rd Qu.: 5154
##  Max.   :288115   Max.   :375006   Max.   :284744   Max.   :29882
##  DUPONT-AIGNAN       HAMON          ASSELINEAU        POUTOU
##  Min.   :   79   Min.   :   217   Min.   :   36   Min.   :   41
##  1st Qu.: 7377   1st Qu.:  8590   1st Qu.: 1227   1st Qu.: 1941
##  Median :13888   Median : 15656   Median : 2238   Median : 3165
##  Mean   :15908   Mean   : 21259   Mean   : 3086   Mean   : 3690
##  3rd Qu.:22620   3rd Qu.: 27188   3rd Qu.: 4070   3rd Qu.: 5046
##  Max.   :65245   Max.   :109550   Max.   :11450   Max.   :13233
##     ARTHAUD        CHEMINADE
##  Min.   :   28   Min.   :   9.0
##  1st Qu.: 1106   1st Qu.: 286.0
##  Median : 1860   Median : 508.5
##  Mean   : 2181   Mean   : 610.1
##  3rd Qu.: 2891   3rd Qu.: 849.5
##  Max.   :10975   Max.   :2338.0
```

2. AFC

Tout d'abord, nous allons effectuer l'analyse factorielle de correspondance(AFC) pour étudier l'association qui existe entre les deux variables : le département et le candidat

L'AFC est une méthode d'analyse de données qui permet d'étudier le liens entre deux variables qualitatives. Basée sur l'inertie, elle consiste à représenter un maximum de l'inertie totale sur le plan factoriel.

C'est une approche géométrique de visualisation des lignes et des colonnes du tableau de contingence en nuage de points à deux dimensions. En effet, elle retourne les coordonnées des éléments des colonnes et des lignes qu'on peut représenter sur un graphique montrant leur association.

Test d'indépendance

Pour étudier l'association entre deux variables, nous supposons qu'il existe une dépendance entre les deux. Nous pouvons évaluer la dépendance des deux variables avec le test de χ^2.

Avec une p-valeur très faible proche de 0, nous pouvons vérifier que la dépendance entre les deux variables est significative.

```
##
##  Pearson's Chi-squared test
##
## data:  donnees_elections
## X-squared = 3630539, df = 1365, p-value < 2.2e-16
```

AFC

L'existence de dépendance est confirmée, nous pouvons donc effectuer l'AFC.

Dans l'analyse factorielle de correspondance, nous regardons l'écart à l'indépendance pour évaluer l'association entre les deux variables.

Pour cela, nous comparons la distribution conditionnelle avec sa distribution marginale du tableau de probabilité obtenue à partir du tableau de contingence. S'il y a une indépendance, la probabilité conditionnelle est égale à la probabilité marginale et s'il n'y a pas d'indépendance, un écart entre les deux est observé.

Nous pouvons calculer l'AFC avec la fonction *CA* du package *FactoMineR*

```
afc = CA(donnees_elections, graph = F)
```

Inertie et Pourcentage d'inertie

Pour avoir une représentation en nuage de points de qualité, nous regardons le pourcentage d'inertie des axes obtenus. Le pourcentage d'inertie d'un axe correspond à l'inertie projetée des lignes ou des colonnes sur l'axe divisé par l'inertie totale des lignes ou des colonnes.

Ici, nous avons un pourcentage d'inertie de 44,24% pour le premier axe qui est énorme et 30,31% pour le deuxième axe qui est également important. En total, les deux premiers axes résument 74,55% de l'écart à l'indépendance. C'est un pourcentage acceptable et nous pouvons effectuer l'interprétaion sur ces deux axes.

```
##          eigenvalue variance.percent cumulative.variance.percent
## Dim.1  3.467861e-02       44.2417275                    44.24173
## Dim.2  2.375968e-02       30.3117455                    74.55347
## Dim.3  8.814431e-03       11.2451353                    85.79861
## Dim.4  5.874345e-03        7.4942799                    93.29289
## Dim.5  2.317789e-03        2.9569528                    96.24984
## Dim.6  1.156770e-03        1.4757663                    97.72561
## Dim.7  7.541410e-04        0.9621061                    98.68771
## Dim.8  3.805197e-04        0.4854535                    99.17317
## Dim.9  3.215832e-04        0.4102643                    99.58343
## Dim.10 1.386597e-04        0.1768971                    99.76033
## Dim.11 9.602619e-05        0.1225068                    99.88284
## Dim.12 8.220477e-05        0.1048739                    99.98771
## Dim.13 9.634225e-06        0.0122910                   100.00000
```

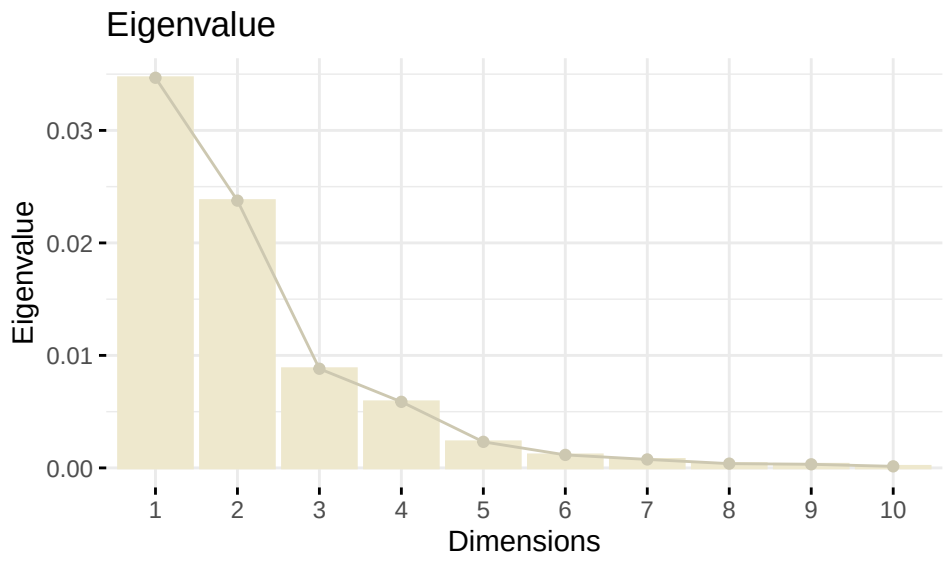
La décroissance des inerties en fonction du rang des axes suggère également le nombre d'axes à conserver. L'inertie de chaque axe donne la quantité d'information retenue par l'axe et l'examination des inerties permet de déterminer le nombre d'axes principaux à considérer dans l'analyse.

La séquence d'inertie peut être représentée avec un graphique en barre. Ici, nous observons que les deux premiers valeurs sont sensiblement grandes que les suivantes. Les deux premiers axes sont prépondérants du point de vue d'inertie donc nous pouvons privilégier ces deux axes dans l'interprétation.

Néanmoins, nous remarquons que l'inertie de chaque axe est très faible.

En l'AFC, l'inertie est comprise entre 0 et 1. Lorsque l'inertie d'un axe est égale à 1, cela signifie qu'il y a une association exclusive entre les modalités des lignes et des colonnes et une force d'opposition de l'axe est très forte. Par exemple, cela correspondrait à un axe qui oppose un départements qui votent 100% un candidat et tous les autres département qui vote 0% ce candidat. C'est une marque d'association très forte entre une modalité d'un variable et une modalité de l'autre.

La valeur d'inertie étant très faibles pour tous les axes, proche de 0 et loins de 1, nous pouvons comprendre que l'association entre les modalités du département et du candidat n'est pas très forte.



Visualisation

Géométriquement, l'association entre les deux variables peut être visualisée par un nuage de points. Le nuage de points montre simultanément les éléments de lignes et de colonnes dans un espace commun. Les lignes sont représentées par des points bleus et les colonnes par des triangles rouges.

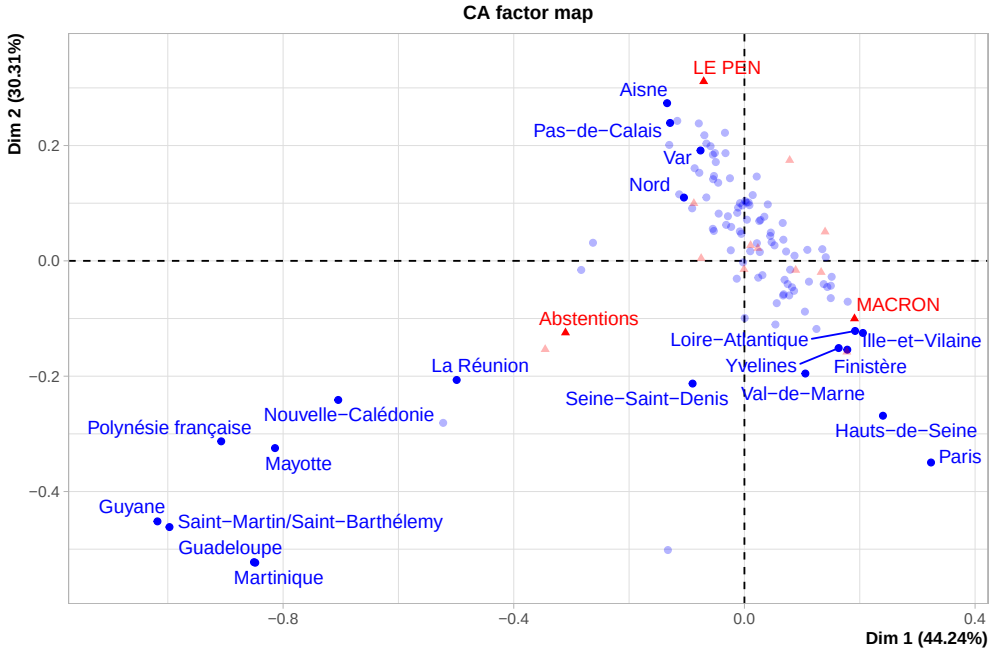
Ici, nous regardons la distance des points par rapport à l'origine des axes, la distance entre les points, et les positions des points.

- La distance des points par rapport à l'origine représente l'inertie du nuage par rapport au centre de gravité. Plus les données s'écartent de l'indépendance, plus les points s'écartent de l'origine.
- La distance entre les points donne une mesure de similitude. Les points de lignes avec un profil similaire sont proches et les points de colonnes avec un profil similaire sont proches sur le graphique.
- Les positions des points expliquent l'opposition des axes et la liaison entre les lignes et les colonnes.

Dans notre jeu de données, nous avons beaucoup de points qui sont proches et se superposent. Nous allons, dans un premier temps, visualiser les points les plus contributifs pour dégager un premier aperçu de l'association entre les variables.

Sur le premier axe, nous observons les départements d'Outre-Mer à proximité d'abstention d'un côté, les departements du Nord-Ouest et la région parisienne de l'Ouest à proximité de Macron de l'autre côté.

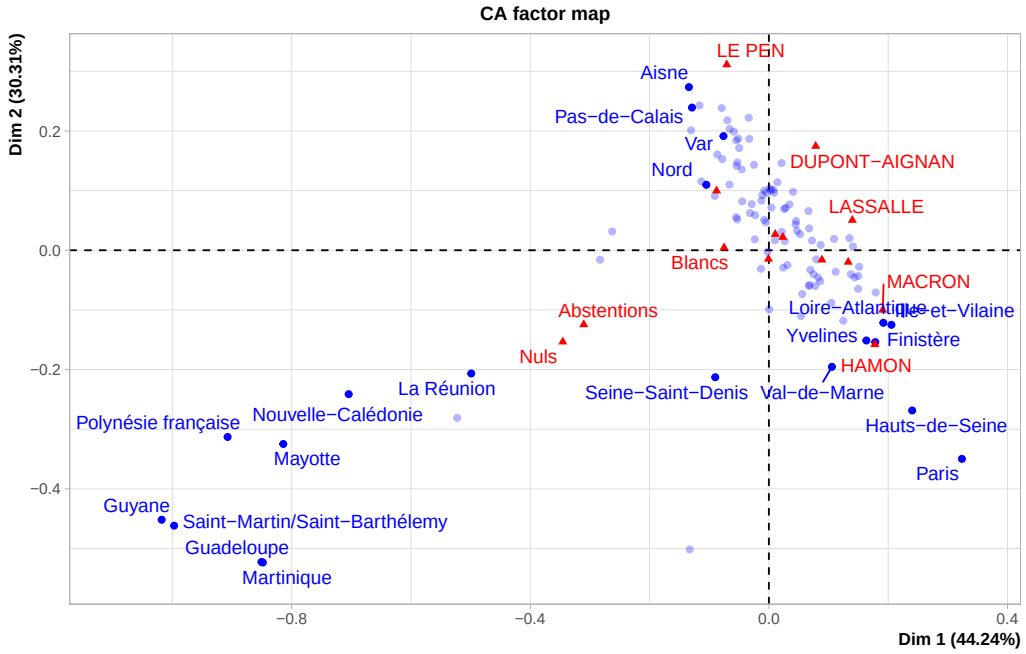
Sur le deuxième axe, nous observons les départements d'Outre-Mer à proximité d'abstention d'un côté, les departements du Nord et du Nord-Est à proximité de LE PEN de l'autre côté.



Affichons maintenant tous les candidats.

Nous remarquons que le premier axe oppose également les candidats masculins avec les candidats non-masculins et le deuxième axe oppose les candidats qui ont exécuté le rôle de ministre dans le passé et qui n'ont jamais été ministre dans le passé.

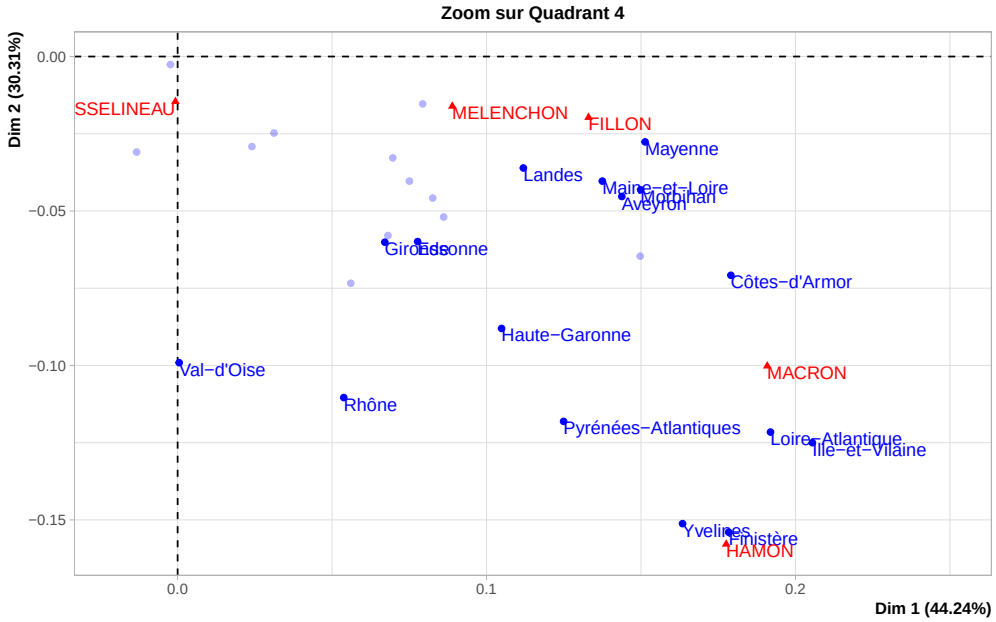
De plus, les candidats qui se trouvent vers la droite sont des candidats qui ne se positionnent pas à l'extrême sur l'échiquier politique et les autres candidats se positionnent à l'extrême sur l'échiquier politique.



Nous pouvons regarder ces éléments du près.

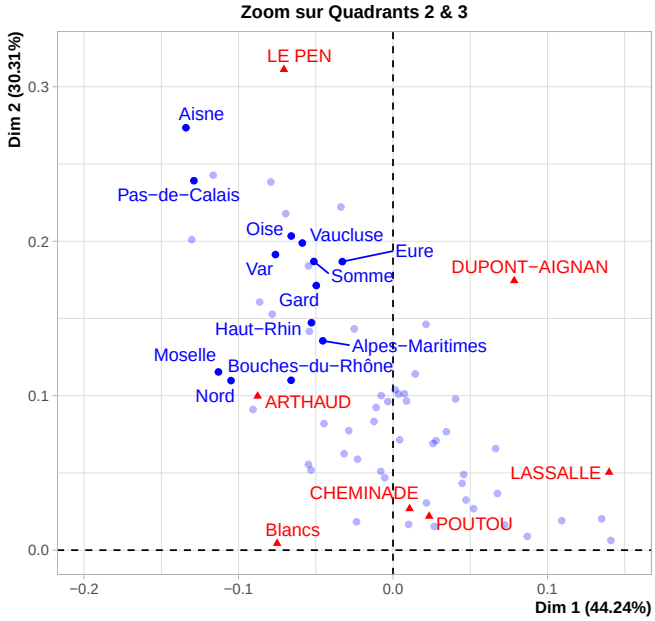
Sur le 4ème quadrant, nous trouvons les départements du Nord-Ouest et du Sud-Ouest avec seuls les candidats qui ont déjà été ministre dans le passé parmi tous les candidats. Nous pouvons supposer une association entre ces départements et ce profil de candidats.

Plus les candidats se trouvent à droite du graphique, moins ils se positionnent extrême sur l'échiquier politique. Nous observons plutôt les département du Nord-Ouest que du Sud-Ouest qui se trouvent prôches des candidats non-extrêmists.



Sur le 2ème quadrant, nous trouvons les départements du Nord-Est et du Sud-Est avec seuls candidats féminins. Nous pouvons supposer une association entre ces départements et les candidats féminins.

Sur les 1er et 2ème quadrants, à part Lassselle qui se trouve à la droite du graphique, les candidats se caractérisent extrêmes qu'ils soient droite ou gauche. Ce sont plus les départments du Nord-Est et du Sud-Est qui se trouvent proches de ces candidats.



Grâce à l'AFC, nous avons pu expliquer l'association entre le département et le candidat. Nous pouvons approfondir notre étude avec la classification sur les candidats.

3. Classification

L'objectif de claissification est d'identifier des classes d'individus similaires dans un jeu de données. Elle consiste à construire des classes d'individus possédant des traits de caractères communs.

Afin d'effectuer une classification sur des données catégorielles, nous avons besoins d'appliquer l'AFC et les transformer en variables quantitatives (axes principaux).

En effet, la classification repose sur la mesure de ressemblance sur les variables quantitatives. Les mesures de ressemblance couramment utilisées sont la distance euclidienne pour la ressemblance entre les individus et le critère de ward pour la ressemblance entre les groupes d'individus.

Nous voudrons avoir les individus d'une même classe proche et les individus de classes différentes éloginées. Cela revient à minimiser l'inertie intra et maximiser l'inertie inter

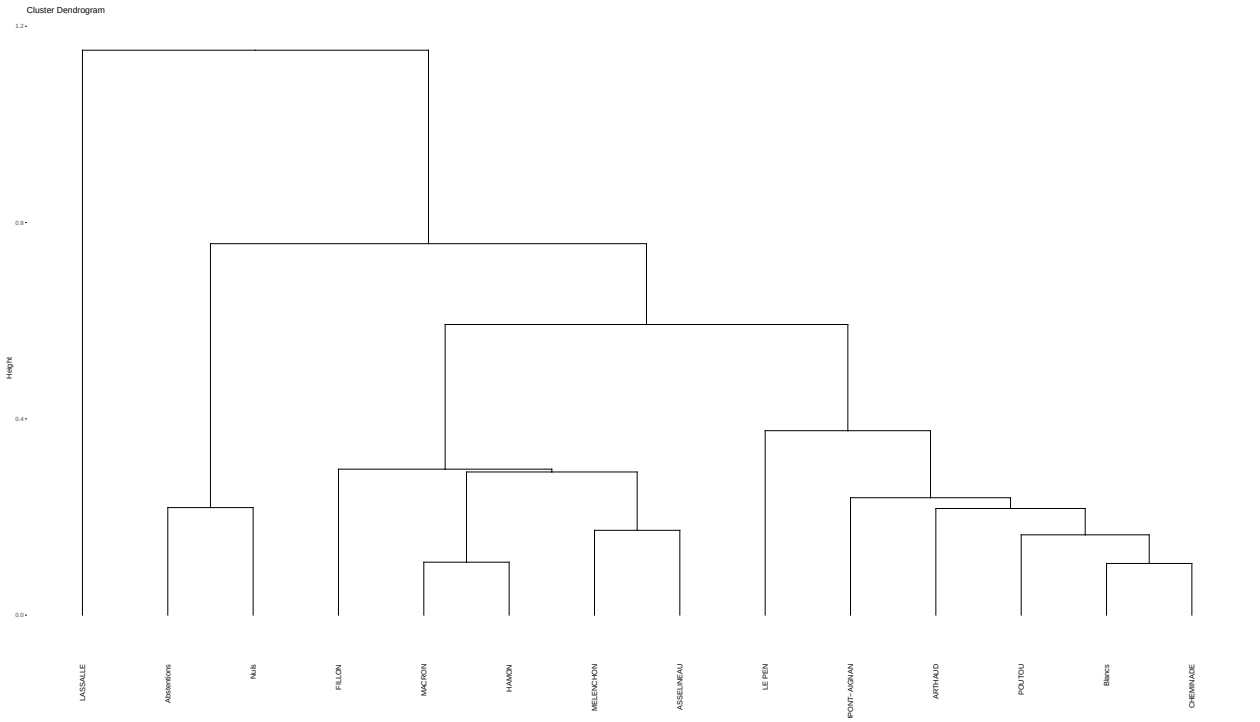
Classification Ascendante Hiérarchique

Nous pouvons reprendre les résultats de AFC obtenu et y appliquer la classification ascendante hiérachique.

La classification ascendante hiérarchique calcule les distances entre chaques points et les deux points les plus proches sont regroupés dans une branche avec une hauteur de branche égale à la distance entre ces deux points. Ces étapes sont itérées et forme un arbre.

À partir de l'arbre, nous pouvons identifier les différentes classes d'observations similaires et detecter le nombre de classes à considérer.

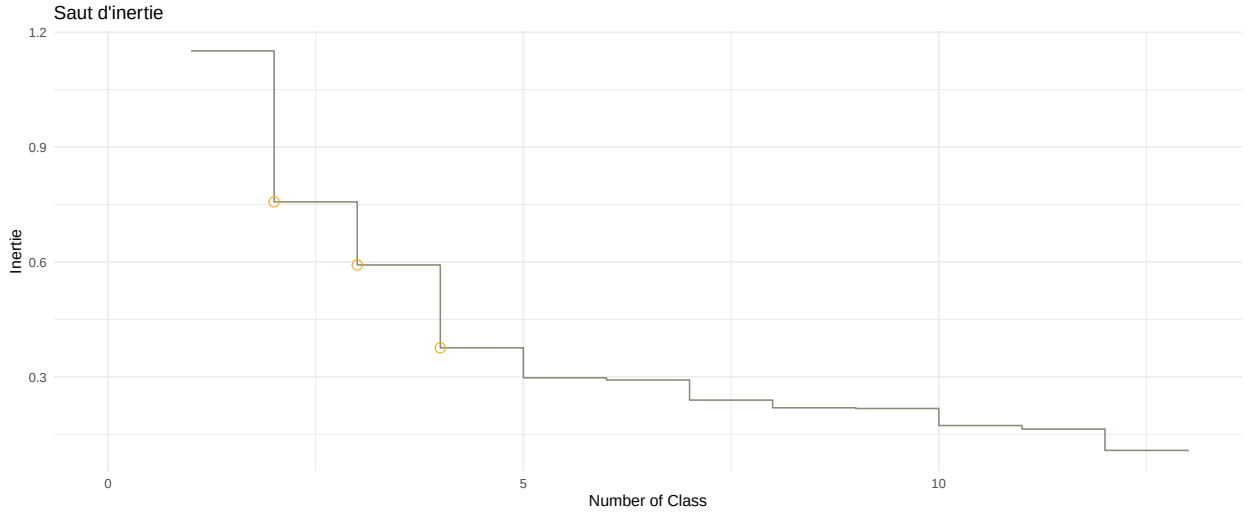
```
distances = dist(afc$col$coord)
cah = hclust(distances, method = "ward.D2")
```



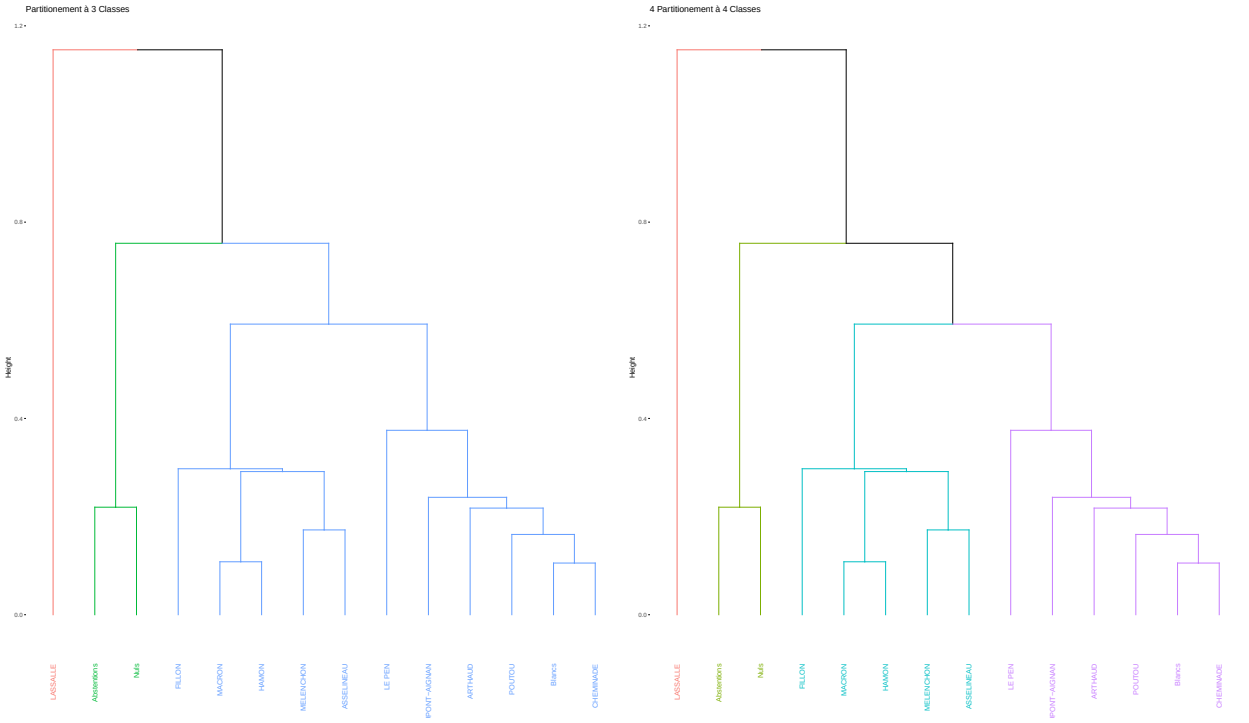
L'arbre ascendante peut être visualisé à l'aide d'un dendrogramme. L'hauteur de chaque branche correspond à la distance entre deux groupes séparés par la branche, donc l'inertie produite. L'hauteur des branches est une première indication qui peut nous guider à choisir le nombre de classes.

Nous avons 3 à 4 branches bien distinctes sur le dendrogramme.

Pour nous aider, nous pouvons représenter les sauts d'inertie de l'arbre selon le nombre de classes retenues. Nous observons un grand saut lorsque le nombre de classe est 2, 3, et 4. Donc une classification de 2 à 4 groupes seraient envisageables.



Nous pouvons mieux visualiser le nombre de partition en ajoutant les couleurs sur le dendrogramme. Les individus de même couleurs appartiennent à la même classe. En passant de 3 classes à 4 classes, nous remarquons la séparation de Fillon, Macron, Hammon, Malenchon, Asselineau avec Le Pen, Dupont-Aignan, Poutou, Blancs, et Cheminade. Néanmoins, dans tous les deux cas, nous avons un seul individu, Lassalle qui se trouve seul dans une classe.



Consolidation

En faite, dans la classification ascendante hiérarchique, nous avons une contrainte d'hiérarchie entre les groupes d'individus, qui n'est pas tout le temps nécessaire. Nous pouvons améliorer la classification obtenue par l'algorithme k-means. Pour cela, nous pouvons appliquer k-means avec le nombre de partitions obtenu lors de la classification ascendante hiérarchique.

L'algorithme de k-means est un algorithme de classification que nous pouvos appliquer avec le nombre de classe à priori. L'algorithme choisit k centres de classe au hasard et affecte tous les points au centre le plus proche. Puis les k centres de gravité sont calculés et ces étapes sont itérés jusqu'à ce que les centres de gravité et les points de classes ne changent plus.

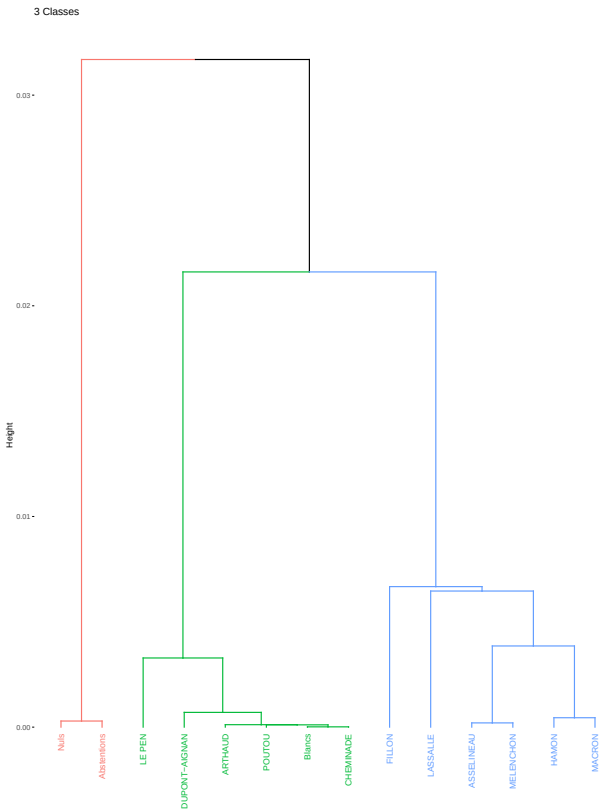
La fonction *HCPC* permet d'effectuer la classification ascéndante hiérarchqiue avec la consolidation par k-means.

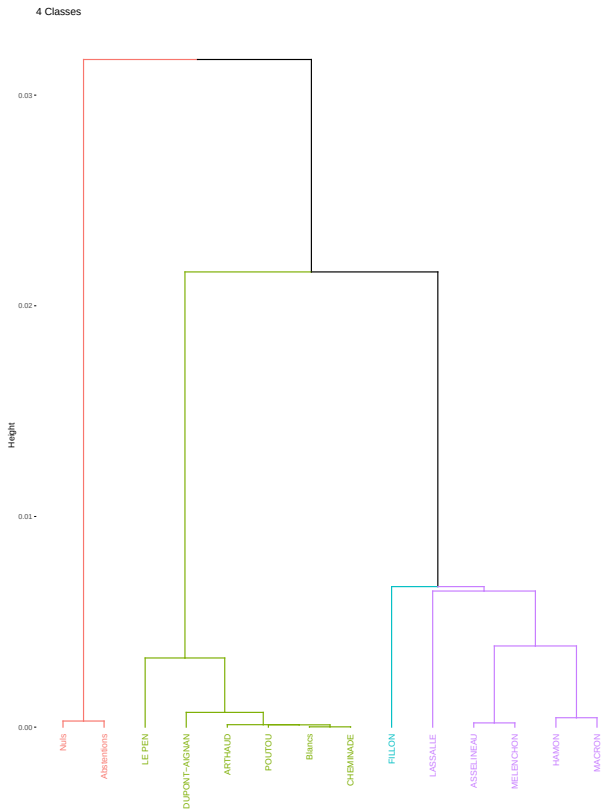
Nous pouvons effectuer HCPC avec 3 et 4 classes.

Par défaut, AFC garde 5 premières dimensions dans le résultat. En reprenant notre résultat obtenu par l'AFC, nous allons considérer seulement les 5 premières dimensions pour la classification. Les 5 premières dimensions contiennent déjà plus de 96% d'inertie donc ce choix est tout à fait correct et nous permet même d'éliminer les bruits contenus dans le reste des dimensions.

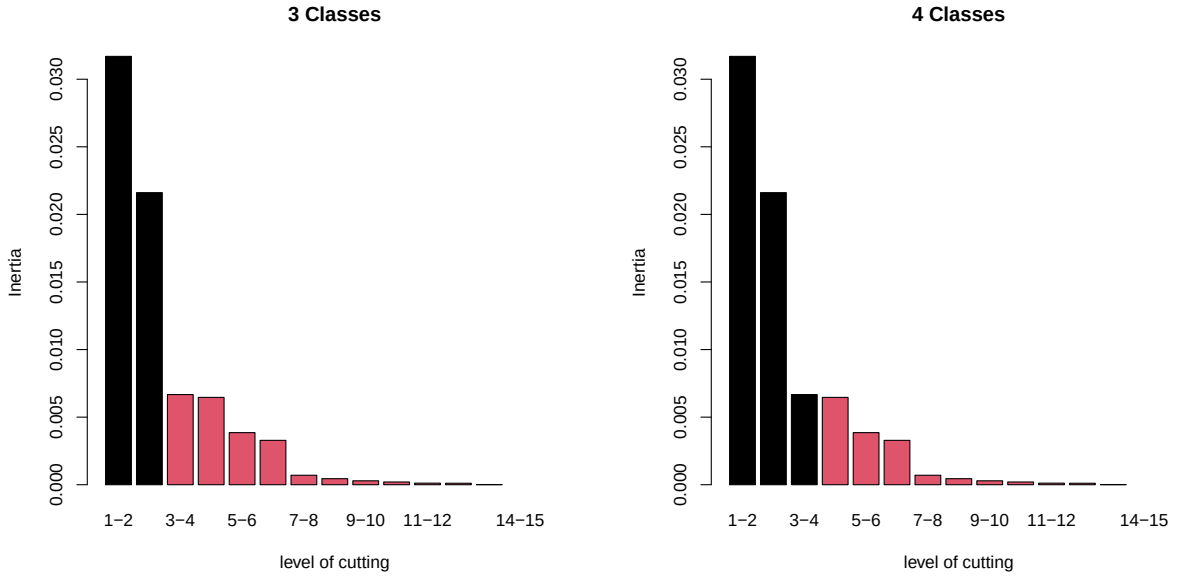
```
hcpc_3 = HCPC(afc, cluster.CA = "columns", nb.clust = 3, consol = F, graph = F)
hcpc_4 = HCPC(afc, cluster.CA = "columns", nb.clust = 4, consol = F, graph = F)
```

Sur le dendrogramme, nous observons 3 branches très distinctes. Le passage de 3 à 4 classes semble moins intéréssant car il sépare un seul individu du reste de la classe.





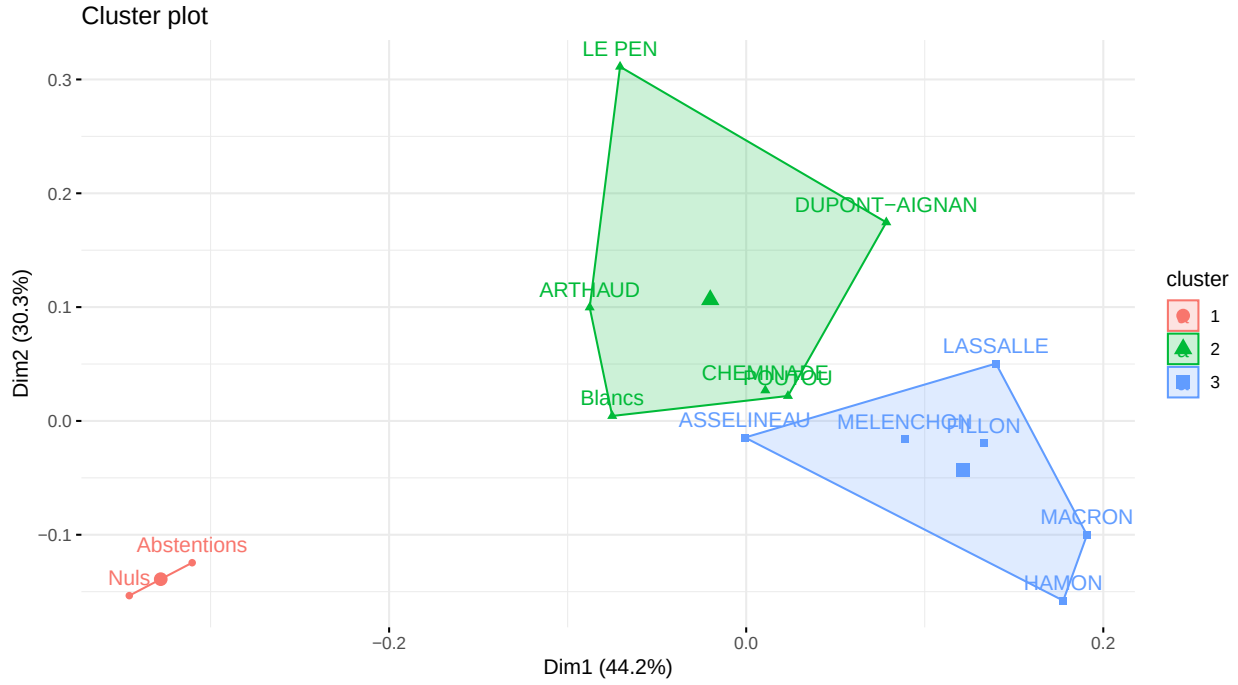
Nous pouvons également nous aider des sauts d'inertie du dendrogramme dans le choix de nombre de classes. Nous observons un grand saut lorsque le nombre de classe est de 3. Le saut d'inertie au 4ème classe est faible. Nous allons donc garder 3 classes.



Nous pouvons visualiser la classification obtenue sur le plan factoriel.

Nous observons trois classes avec leur centre de gravité et nous voyons que les individus sont affectés par la classe ayant le centre de gravité le plus proche d'eux.

La classification sépare les votes nuls et absents dans la classe 1, les candidats avec les caractères politiques extrêmes dans la classe 2, et les candidats qui se présentent pour la première fois à l'élection présidentielle (sauf Melenchon) et qui étaient ex-ministères (sauf Lassalle) dans la 3ème classe.



Caractérisation des classes

Pour mieux comprendre les caractéristiques de classes, nous pouvons regarder les individus représentatifs de chaque classe, qui sont les indivius proches du centre de gravité de la classe.

Dans la classe 1 correspond à l'absence de choix de candidat. La classe 2 est représentée plutôt par les politiciens extrêmes et la classe 3 par les ex-ministres.

```
## Cluster: 1
## Abstentions        Nuls
##   0.1095613   0.1095613
## ------------------------------------------------------------
## Cluster: 2
##     CHEMINADE        Blancs DUPONT-AIGNAN       ARTHAUD        POUTOU
##     0.1043582     0.1178473     0.1329347     0.1414009     0.1431068
## ------------------------------------------------------------
## Cluster: 3
##     MACRON  MELENCHON      HAMON ASSELINEAU     FILLON
##  0.1684662  0.1836588  0.1912463  0.2224281  0.2458822
```

Nous pouvons également regarder les individus les plus caractéristiques de la classe dans le sens où ils sont les plus éloignés de tous les autres classes.

Dans la classe 1 c'est le vote nul qui la caractérise le plus, dans la classe 2 Le Pen est l'individu le plus caractérisant de la classe, et Lassalle dans la classe 3.

```
## Cluster: 1
##        Nuls Abstentions
##   0.4303722   0.3898396
## ------------------------------------------------------------
## Cluster: 2
##        LE PEN       ARTHAUD DUPONT-AIGNAN        Blancs        POUTOU
##     0.4241120     0.3734274     0.3087398     0.2536986     0.2209380
## ------------------------------------------------------------
## Cluster: 3
##  LASSALLE     HAMON    FILLON    MACRON MELENCHON
## 0.8490761 0.3442827 0.3129342 0.3080157 0.2615573
```

4. Conclusion

Nous avons effectué une analyse de données avec l'analyse factorielle de correspondance et la classification.

L'enchaînement de l'AFC et la classification a un grand avantage d'éliminer les bruits que contiennent les dernières composantes. Nous éliminons ainsi l'aléatoire avant de faire la classification qui permet d'obtenir une classification plus stable où les classes seront moins affectées par l'ajout ou le retrait de quelques individus.

La visualistaion en 3D du dendrogramme sur le plans AFC nous donne une bonne visualisation de notre étude. Elle nous donne une représentation synthétique de notre analyse dans la globalité donnant l'information sur le plan factoriel, l'arbre hiérarchique et la classification.

Hierarchical clustering on the factor map