Université Gustave Eiffel

Master Actuariat

Analyse de données et SAS

# Projet : Mettre en évidence des liens entre géographie et profil politique et Déterminer le prix d'un bien immobilier

***Élèves :***
Marie Meyer
Linh Chi

***Enseignant :***
Thomas Bonis

Année académique 2023-2024

# Table des matières

# Introduction

Dans ce rapport nous allons nous allons dans un premier temps réaliser une analyse de données sur les résultats du premier tour des élections présidentielles au niveau de la France pour ainsi chercher à mettre en évidence des liens entre géographie et profl politique (vote). Nous explorerons dans une seconde partie la régression du prix d'un bien immobilier à l'aide de divers descripteurs (taille, nombre de pièce, jardin, localisation, etc. )

Les objectifs sont donc de **mettre en évidence des liens entre géographie et profil politique** et de **déterminer le prix d'un bien immobilier**.

# 1 Exercice 1 : (Analyse des profils de bureaux de vote)

Dans ce premier exercice, nous réalisons une analyse de données sur les résultats du premier tour des élections présidentielles au niveau de la France. Comme dit pécédemment, nous cherchons à mettre en évidence des liens entre géographie et profil politique (vote).

## 1.1 Question 1

```
proc import datafile='/home/Marie/DM2/resultats.xlsx'
            out=votlib.resultats
            dbms=xlsx
            replace;
    getnames=yes;
    delimiter=';';
run;


data resultats;
run;
```

Nous avons dans un premier temps importer l'ensemble de la base de données sur SAS et sur R.

Nous obtenons un tableau composé de 69 682 lignes et de 105 colonnes. Ce jeu de données décrit 69 962 bureaux de votes qui ont voté par 105 variables qui correspondent à différents profil de votes tels que l'absentions, les votes nuls...
Les 69 962 individus vivent dans $\mathbf{R}^{105}$ et les 105 variables vivent dans $\mathbf{R}^{69962}$.

## 1.2 Question 2

Nous voulons maintenant **réfléchir aux descripteurs pertinents pour dresser un profil politique (vote) de chaque bureau de vote**.

Il y a plusieurs variable qui ne peuvent pas être retenue commme variables active d'une Analyse en Composantes Principales car elles ne sont pas quantitatives.

Nous avons sélectionnés 14 variables qui nous semblent pertinentes : Libellé de la commune, Abstention et le pourcentage de voix par rapport aux exprimés de chaque candidat (soit 12 variables correspondantes) à l'aide du code suivant :

```
data resultats (drop = "Libelle de la co"n NPanneau Nom Sexe Prenom ... DA
   ="%Voix/Exp12"n));
run;
```

Ainsi, dans notre nouveau jeu de données, nous avons :
- Nathalie Arthaud, extrême gauche
- Fabien Roussel, extrême gauche
- Emmanuel Macron
- Jean Lassalle, droite
- Marine Le Pen, extrême droite
- Eric Zemmour, extrême droite

- Jean-Luc Mélenchon, extrême gauche
- Anne Hidalgo, gauche
- Yannick Jadot, gauche
- Valérie Pécresse, droite
- Philippe Poutou, extrême gauche
- Nicolas Dupont-Aignant, extrême droite

**Discutons maintenant d'une éventuelle normalisation de ces descripteurs.**

Il est préférable de faire une ACP normée qu'une ACP non normée car dans l'ACP, les variables initiales peuvent être mesurées dans des échelles différentes. les variances des variables ne sont en effet pas toutes de même ordre de grandeur : il faut réduire les données pour donner le même poids à toutes les variables dans le calcul de la distance entre deux individus. De fait, il peut y avoir des variables qui en dominent d'autres (voire une seule variable qui domine tout le reste).

Ainsi, normaliser les variables permet d'avoir une importance égale pour toutes.

## 1.3 Question 3

Nous voulons dans cette question **réaliser une ACP**. Comme dit précedemment, nous allons réaliser une ACP normée.

Nous ne gardons que les colonnes de pourcentage des candidats par rapport aux inscrits et par rapport aux départements.

Pour bien comprendre l'analyse des composantes principales (ACP), commençons par faire une analyse rapide de nos données.

| Variable | Libellé | N | Moyenne | Ec-type | Minimum | Maximum |
|---|---|---|---|---|---|---|
| Abstention | % Abs/Ins | 69682 | 24.1192173 | 9.9624689 | 0 | 100.0000000 |
| Arthaud | % Voix/Ins | 69682 | 0.4470951 | 0.4628226 | 0 | 10.0000000 |
| Roussel | AH | 69682 | 1.7455310 | 1.3318652 | 0 | 33.3300000 |
| Macron | AO | 69682 | 19.6282373 | 6.7647284 | 0 | 56.5200000 |
| Lassalle | AV | 69682 | 2.9892682 | 2.9187652 | 0 | 56.0600000 |
| LePen | BC | 69682 | 18.9107336 | 8.2347934 | 0 | 75.0000000 |
| Zemmour | BJ | 69682 | 5.1593714 | 2.6837482 | 0 | 50.0000000 |
| Melenchon | BQ | 69682 | 14.8884609 | 7.1021386 | 0 | 100.0000000 |
| Hidalgo | BX | 69682 | 1.2790844 | 1.0133527 | 0 | 35.4800000 |
| Jadot | CE | 69682 | 3.1029412 | 2.0049614 | 0 | 100.0000000 |
| Pécresse | CL | 69682 | 3.6772538 | 2.3354254 | 0 | 49.4300000 |
| Poutou | CS | 69682 | 0.6047400 | 0.5638601 | 0 | 15.4900000 |
| Dupont-Aignant | CZ | 69682 | 1.6748358 | 1.1863351 | 0 | 53.3300000 |

D'après la sortie de notre 'proc means', nous avons dans un premier temps accès à des informations sur chacune des variables. Par exemple, à l'issue du premier tour de l'élection présidentielle de 2022, l'abstention se situe à **24 %** des inscrits sur les listes électorales (soit 12,8 millions de personnes selon le journal Le Monde). Emmanuelle Macron fait une moyenne de **19.63 %** toutes villes confondues, qui représente le taux le plus élevés des candidats. Au contraire, la candidate Nathalie Arthaud est celle qui a le plus mauvais score, avec un taux de **0.45 %**.

Maintenant, pour réaliser notre ACP, nous utilisons la procédure 'proc princomp' avec ncomp=4 (cf 1.3).

```
proc princomp data=resultats out=votlib.resultats n=5 PLOTS(ncomp=4) =(SCORE
    pattern(circle vector));
id "Code du departement"n;
var Arthaud Roussel Macron lassalle LePen Zemmour Melenchon Hidalgo Jadot
   Pecresse Poutou "Dupont-Aignant"n;
run;
```

Nous obtenons les résultats suivant :

La procédure PRINCOMP

| | |
|---|---|
| Observations | 69682 |
| Variables | 13 |

Statistiques simples

| | Abstention | Arthaud | Roussel | Macron | Lassalle | LePen | Zemmour | Melenchon | Hidalgo | Jadot | Pécresse | Poutou | Dupont-Aignant |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Moyenne | 24.11921730 | 0.4470950891 | 1.745530984 | 19.62823728 | 2.989268247 | 18.91073362 | 5.159371430 | 14.88846087 | 1.279084412 | 3.102941219 | 3.677253810 | 0.6047399615 | 1.674835826 |
| StD | 9.96246885 | 0.4628225809 | 1.331865206 | 6.76472842 | 2.918765160 | 8.23479341 | 2.683748212 | 7.10213855 | 1.013352664 | 2.004961371 | 2.335425381 | 0.5638601199 | 1.186335086 |

FIGURE 1 – Statistiques simples

Matrice de corrélation

| | | Abstention | Arthaud | Roussel | Macron | Lassalle | LePen | Zemmour | Melenchon | Hidalgo | Jadot | Pécresse | Poutou | Dupont-Aignant |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Abstention | % Abs/Ins | 1.0000 | -.0623 | -.2221 | -.5298 | -.3099 | -.3155 | -.3326 | 0.0939 | -.2294 | -.3525 | -.3564 | -.1508 | -.2572 |
| Arthaud | % Voix/Ins | -.0623 | 1.0000 | 0.0587 | -.0617 | 0.0191 | 0.1587 | -.1133 | -.0607 | -.0038 | -.0902 | -.0561 | 0.0851 | 0.0701 |
| Roussel | AH | -.2221 | 0.0587 | 1.0000 | -.0472 | 0.1634 | 0.0744 | -.0604 | 0.0338 | 0.1436 | 0.0086 | -.0770 | 0.0796 | -.0209 |
| Macron | AO | -.5298 | -.0617 | -.0472 | 1.0000 | -.0523 | -.2170 | 0.2021 | -.2559 | 0.1237 | 0.4422 | 0.3979 | -.0323 | 0.0264 |
| Lassalle | AV | -.3099 | 0.0191 | 0.1634 | -.0523 | 1.0000 | 0.0932 | 0.0218 | -.1527 | 0.2044 | -.0633 | 0.0796 | 0.1385 | 0.1060 |
| LePen | BC | -.3155 | 0.1587 | 0.0744 | -.2170 | 0.0932 | 1.0000 | 0.0929 | -.5181 | -.1423 | -.3693 | -.0522 | 0.0601 | 0.2637 |
| Zemmour | BJ | -.3326 | -.1133 | -.0604 | 0.2021 | 0.0218 | 0.0929 | 1.0000 | -.3007 | -.0805 | 0.0921 | 0.2866 | -.0738 | 0.1068 |
| Melenchon | BQ | 0.0939 | -.0607 | 0.0338 | -.2559 | -.1527 | -.5181 | -.3007 | 1.0000 | 0.0957 | 0.1545 | -.3171 | 0.0221 | -.2520 |
| Hidalgo | BX | -.2294 | -.0038 | 0.1436 | 0.1237 | 0.2044 | -.1423 | -.0805 | 0.0957 | 1.0000 | 0.2041 | -.0135 | 0.0622 | -.0394 |
| Jadot | CE | -.3525 | -.0902 | 0.0086 | 0.4422 | -.0633 | -.3693 | 0.0921 | 0.1545 | 0.2041 | 1.0000 | 0.1422 | 0.0549 | -.0246 |
| Pécresse | CL | -.3564 | -.0561 | -.0770 | 0.3979 | 0.0796 | -.0522 | 0.2866 | -.3171 | -.0135 | 0.1422 | 1.0000 | -.0418 | 0.1065 |
| Poutou | CS | -.1508 | 0.0851 | 0.0796 | -.0323 | 0.1385 | 0.0601 | -.0738 | 0.0221 | 0.0622 | 0.0549 | -.0418 | 1.0000 | 0.0463 |
| Dupont-Aignant | CZ | -.2572 | 0.0701 | -.0209 | 0.0264 | 0.1060 | 0.2637 | 0.1068 | -.2520 | -.0394 | -.0246 | 0.1065 | 0.0463 | 1.0000 |

FIGURE 2 – Matrice de corrélations

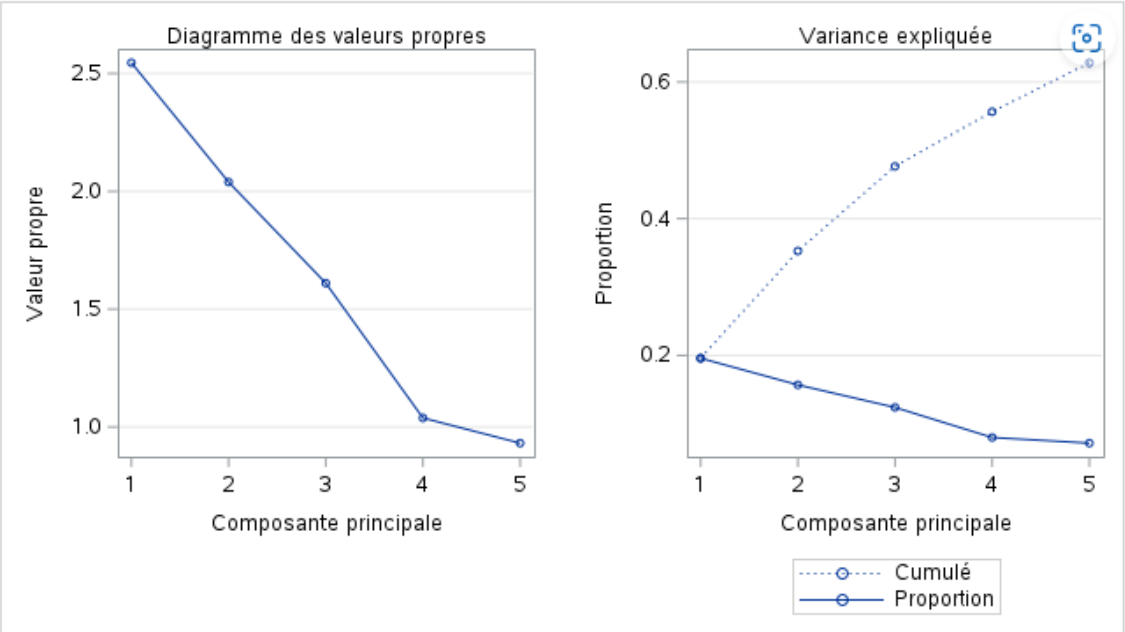
Valeurs propres de la matrice de corrélation

| | Valeur propre | Différence | Proportion | Cumulé |
|---|---|---|---|---|
| 1 | 2.54538066 | 0.50606866 | 0.1958 | 0.1958 |
| 2 | 2.03931200 | 0.42897615 | 0.1569 | 0.3527 |
| 3 | 1.61033585 | 0.57209235 | 0.1239 | 0.4765 |
| 4 | 1.03824350 | 0.10679714 | 0.0799 | 0.5564 |
| 5 | 0.93144637 | | 0.0716 | 0.6281 |

FIGURE 3 – Valeurs propres de la matrice de corrélation

| Vecteurs propres | | | | | | |
|---|---|---|---|---|---|---|
| | | Prin1 | Prin2 | Prin3 | Prin4 | Prin5 |
| Abstention | % Abs/Ins | -.517719 | -.040162 | -.247372 | -.061537 | 0.060270 |
| Arthaud | % Voix/Ins | -.002922 | -.191071 | 0.214770 | 0.651309 | -.423190 |
| Roussel | AH | 0.060446 | -.026878 | 0.455066 | -.228655 | -.462789 |
| Macron | AO | 0.425234 | 0.305377 | -.151002 | 0.176659 | -.141625 |
| Lassalle | AV | 0.187395 | -.130803 | 0.402669 | -.424678 | 0.235042 |
| LePen | BC | 0.148286 | -.571872 | 0.073281 | 0.030851 | -.096055 |
| Zemmour | BJ | 0.332623 | -.055094 | -.284564 | -.221111 | 0.035788 |
| Melenchon | BQ | -.319797 | 0.390369 | 0.214692 | 0.060656 | 0.069116 |
| Hidalgo | BX | 0.114486 | 0.237338 | 0.413189 | -.202923 | -.099285 |
| Jadot | CE | 0.234656 | 0.484174 | 0.045280 | 0.223341 | 0.032049 |
| Pécresse | CL | 0.392732 | 0.064389 | -.249140 | -.050289 | 0.021222 |
| Poutou | CS | 0.055831 | -.034427 | 0.364297 | 0.340987 | 0.662116 |
| Dupont-Aignant | CZ | 0.227907 | -.268784 | 0.025558 | 0.211629 | 0.250570 |

FIGURE 4 – Vecteurs propres



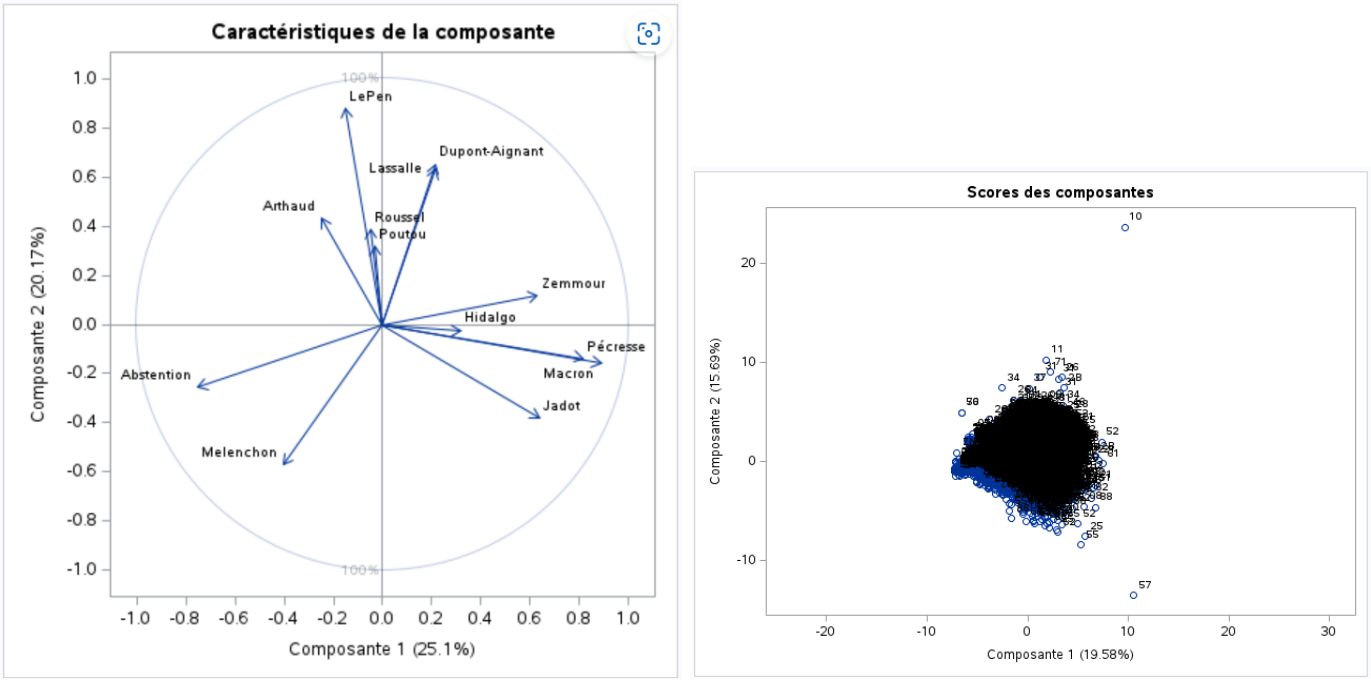

Nous souhaitons désormais voir les corrélations entre les différents candidats. Nous observons pour cela la matrice des corrélations (Figure 2).

Pour analyser cette matrice, nous comparons la corrélation des variables deux à deux. Une corrélation proche de 0 signifie que les deux variables ne sont pas corrélées et une corrélation proche de 1 ou de -1 signifie que les deux variables sont très corrélées.

Par exemple, on voit que la corrélation entre Yannick Jadot et Emmanuel Macron est de **0.44**. Le fait que ce nombre soit positif signifie que de manière général, si Yannick Jadot fait un bon score dans une ville alors Emmanuel Macron fera également un bon score dans cette ville et vis et versa.

Nous voulons en savoir plus sur les corrélations des candidats. De plus, nous voulons associer les différentes villes à des candidats. Pour cela, nous allons analyser les résultats de notre ACP.

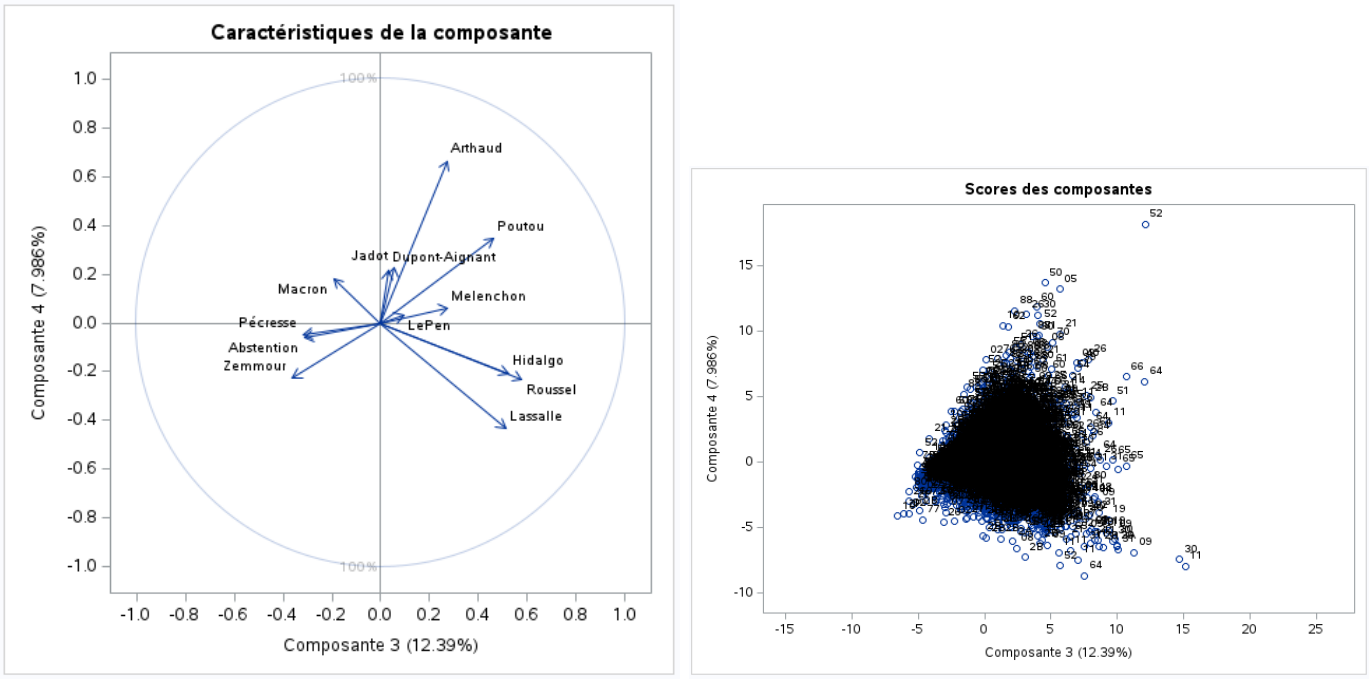
Nous obtenons la représentations graphique des variables projetées sur le 1e plan factoriel (cercle des corrélations) suivant :




(a) Cercle des corrélations (b) Nuage des points

FIGURE 5 – Représentation des individus et des variables sur le 1e plan factoriel

Sur le 2e plan factoriel, nous avons :




(a) Cercle des corrélations (b) Nuage des points

FIGURE 6 – Représentation des individus et des variables sur le 2e plan factoriel

Sur les nuages de points, nous ne voyons pas grand chose à cause du nombre important des bureaux de votes analysées.

Sur les graphiques précedent (Figures 5 et 8), nous avons aplatit toutes les variables sur 4 dimensions (nous allons donc perdre énormément d'information). Chaque axe va représenter un certain pourcentage de l'information. Nous pouvons de plus voir que le 1e axe des abscisses (1e axe) représente **19.58 %** de l'information et celui des ordonnés (2e axe) **15.69 %** ; et le 2e axe des abscisses (3e axe) représente **12.39 %** et celui des ordonnée (4e axe) **7.986 %**. Ainsi, les graphiques représentent **55.646 %** de l'information. C'est peu mais c'est bien souvent le cas avec des données réelles.

Pour augmenter la qualité de la représentation, nous pouvons supprimer des variables très corrélées où des villes possèdent des scores spéciaux. Dans le cadre de ce DM, nous avons décider de ne rien supprimer et de se contenter de cette qualité. Mais en analysant le graphique, nous aurions pu supprimer les villes Troyes, Molring et 57 qui semblent avoir des scores distincts des autres villes.

Une fois s'être fait l'idée de la qualité de la représentation, nous allons tenter de définir les axes avec ce dernier.

**Combien d'axes semblent pertinents (justifiez votre choix) ?**

Pour savoir combien d'axes nous retenons pour l'analyse, nous utilisons le critère de Kaiser. Pour cela, nous avons besoin de connaître les valeurs propres.

D'après la Figure 3, nous obtenons 5 valeurs propres.

Nous avons :
$\lambda_1 = 2{,}545 > 1$
$\lambda_2 = 2{,}039 > 1$
$\lambda_3 = 1{,}610 > 1$
$\lambda_4 = 1{,}038 > 1$
$\lambda_5 = 0{,}931 < 1$

Selon le critère de Kaiser, nous ne retenons uniquement les composantes qui présentent une valeur propre supérieure à 1, donc que 4 axes car $\lambda_1$, $\lambda_2$, $\lambda_3$, $\lambda_4 > 1$ alors que $\lambda_5 < 1$.
De plus, 4 étant un nombre pair, 4 axes semble en effet acceptable pour les visualitations.

Pour avoir le pourcentage d'inertie expliquée par ces 4 composantes, nous faisons :

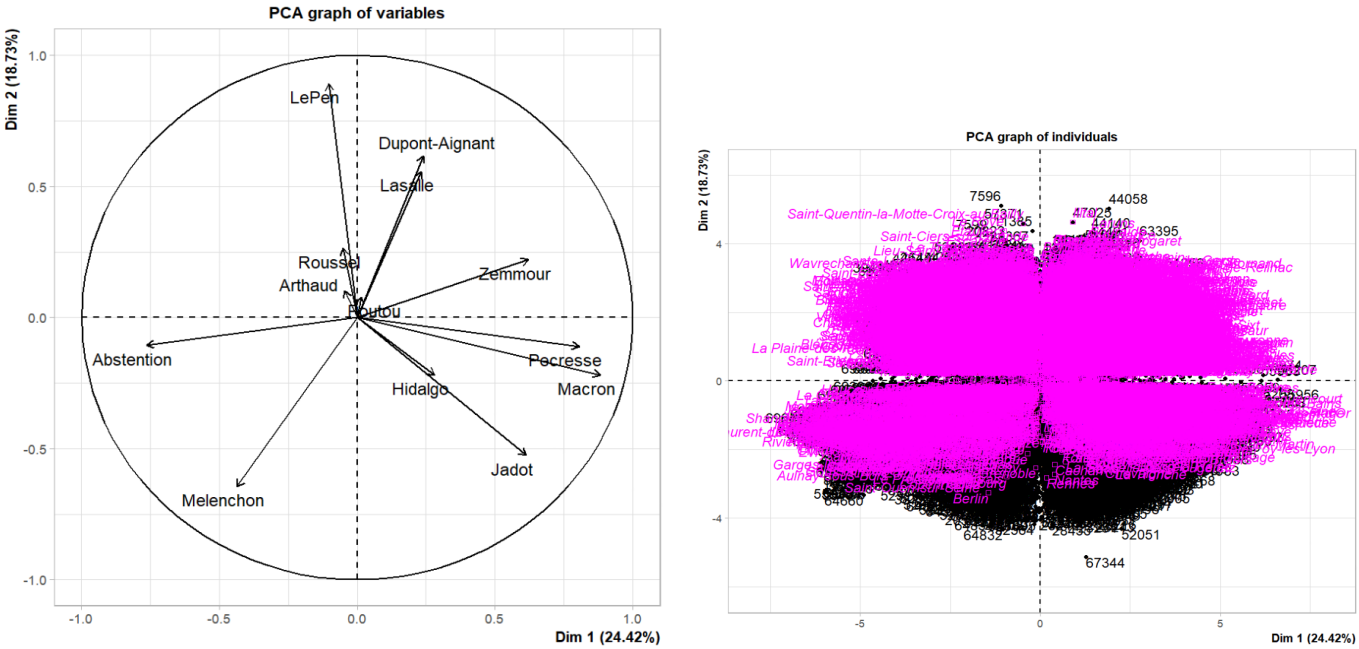
$$I_{explique} = \frac{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4}{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + \lambda_5} = 0.8859 = 88.59\%$$

La projection du nuage sur le 1e plan factoriel retranscrit **88.59 %** de l'inertie (dispersion) du nuage. Ainsi, environ **88.59 %** de l'inertie est expliquée par les 4 premiers axes.

Or, en choisissant 2 axes, nous avons une inertie de 56,15 % . L'inertie est moins bonne mais elle reste tout de même supèrieur à 50 %. De plus, pour une analyse plus approfondie, c'est plus simple de choisir 2 axes.

Nous allons donc dans les étapes suivantes choisir 2 axes et sélectionner seulement les communes avec plus de 1000 inscrits pour ne pas faire planter SAS et R.

Nous avons donc le nouveau cercle des corrélations et le nouveau nuage des points suivant (nous avons réaliser une AC sur SAS et sur R et obtenons les mêmes sorties, ce qui permet de valider nos résultats) :




(a) Cercle des corrélations (b) Nuage des points

FIGURE 7 – Représentation des individus et des variables sur le 1e plan factoriel (en prenant les bureaux de votes ayant plus de 1000 inscrits)

Le 1e axe représente ici **24.42 %** de l'information et le 2e axe **18.73 %**.

## 1.4 Question 4

Pour le nuage des variables/ cercle des corrélations (voir Figure 7) , nous voulons **regarder pour chaque axe les variables qui contribuent le plus à l'inertie de l'axe, pour les côtés positifs et négatifs.**

Sur le cercle des corrélations, nous y retrouvons l'ensemble des variables (des candidats dans notre cas). Chaque variable est représentée par une flèche et la longueur des flèches va montrer si une variable est bien représentée ou non dans ce graphique. Une flèche proche du cercle des corrélations indique une bonne représentation de la variable, à l'inverse une flèche proche du centre indique une variable mal représentée. Nous observons ainsi que les candidats Le Pen, Abstention, Macron, Jadot, Mélenchon sont bien représentées sur le 1e plan factoriel.

De surcroît, deux flèches proches signifient qu'elles sont corrélées (corrélation proche de 1). Deux flèches opposées signifient qu'elles sont également corrélées (corrélation proche de -1). A l'inverse deux flèches orthogonales signifie qu'il n'y a aucune corrélation entre les variables. Enfin, il est impossible de déduire quelque chose de flèches ne se rapprochant pas du cercle.

Nous pouvons voir sur notre graphique une certaine corrélation entre **Zemmour** et l'**Abstention** car nous voyons une tendance qui dit que plus il y a d'Abstention, moins il y a de vote pour Zemmour et inversement. Nous pouvons aussi voir que les candidats **Dupont-Aignant** et **Lassalle** ont une certaine corrélation (Si l'un fait un gros score dans une ville, alors l'autre fera aussi un gros score et inversement).

Enfin, le cercle des corrélations définit donc les axes. Nous retrouvons ces axes sur le nuage de points. Pour interpréter ce nuage de points, il faut imaginer une superposition entre ce dernier et le cercle des corrélations.

Les coordonnées des points-variables sur l'axe $\alpha$ correspondent à la corrélation entre les variables et la composante principale $C_\alpha$.

Pour interpréter l'axe factoriel $\alpha$, nous allons donc regarder les variables ayant une grande coordonnée (en valeur absolue).

Axe 1 : il oppose du côté positif les candidats Macron, Pécresse, Le Pen, Zemmour, Jadot et du côté négatif le candidat Mélenchon et l'Abstention.
Ainsi, les candidats (variables) qui contribuent le plus à l'inertie de l'axe sont les variables proches de l'axes et proches du cercle des corrélations, soient les candidats Macron, Pecresse et la variable Abstention.

Axe 2 : il oppose du côté positif les candidats Le Pen, Dupont-Aignant, Lassalle, Zemmour et du côté négatif les candidats Mélenchon, Jadot, Macron.
Ainsi, les candidats (variables) qui contribuent le plus à l'inertie de l'axe sont les variables proches de l'axes et proches du cercle des corrélations, soient le candidat Le Pen.

Comme dit précédemment, le premier axe représente **24.42 %** de l'information. Si nous observons le cercle des corrélations, nous remarquons les bureaux de votes à gauche du nuage de points ont tendance à s'absenter. Cette axe pourrait représenter l'Abstention.

Le second axe représente **18.42 %** de l'information. Si on étiquette les candidats en fonction de leur orientation gauche/droite, nous nous aperçevont que le deuxième axe pourrait décrire cette orientation. Ainsi, les villes en haut du nuage de points ont tendance à voter pour des candidats de droite, et les villes en bas de ce nuage de points votent pour des partis de droite.

**Quelles sont, toujours pour chaque axe, les variables bien représentées sur l'axe ?**

Comme dit précèdemment, les variables qui sont bien représentés sont les variables qui sont proches du cercle des corrélations.

Axe 1 : Les candidats (variables) bien représentés par cette axe sont les candidats Pecresse, Macron, Zemmour et l'Abstention.

Axe 2 : Les candidats (variables) bien représentés par cette axe sont les candidats Le Pen, Dupont-Aignant, Lassalle.

## 1.5 Question 5

Nous voulons maintenant **déterminer les bureaux de votes qui contribuent le plus à chaque axe, puis décrire ces bureaux (d'un point de vue géographique).**

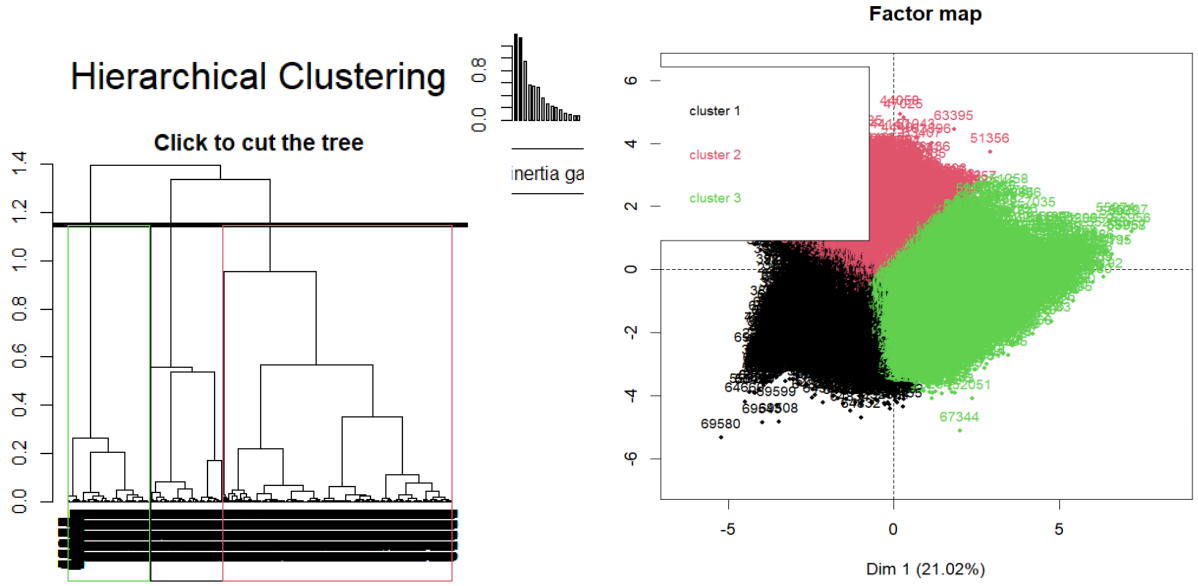
Pour cela, nous devons étudier le 'score des composantes' de notre ACP. Or, nous remarquons que celui-ci est illisible.

Nous allons donc réaliser une classification hiérarchique en appliquant les résultats de notre ACP.

Les deux individus (villes) les plus proches (ayant des résultats proches) sont regroupés et forment un unique individu ayant les caractéristiques moyennes des deux premiers individus. Et ainsi de suite jusqu'à atteindre un unique groupe.

La Figure 8 présente cette algorithme. La taille des branches dépend de la distance des deux groupes regroupés. De plus, l'algorithme nous propose un nombre de classe optimal en traçant un trait horizontal. Ici par exemple, il nous est proposé de partitionner les villes en 3 classes. L'inertie est également affiché en haut à droite permettant de justifier la coupe de l'arbre.

Nous choisissons donc de suivre la proposition de coupe et de sélectionner 3 classes. Nous obtenons donc le nuage de points coloré en fonction de l'appartenance des classes :




(a) Dendogramme (b) Clusther

FIGURE 8 – Classification des bureaux de votes de plus de 1000 inscrits

Nous pouvons faire une première analyse des classes de manière visuelle. Si nous nous tenons à ce que nous avons déduit de l'analyse du cercle des corrélations, la **classe 1** correspond aux villes ayant un fort taux d'abstention et qui votent le plus à l'extrême gauche (Mélenchon), la **classe 3** aux villes qui votent le plus à gauche et la **classe 2** aux villes qui votent le plus à droite.

Maintenant, analysons les caractéristiques des groupes. La commande suivante, nous donne des informations sur les groupes :

```
vote.clust$desc.var
```

Nous obtenons la liste suivante :

$`1`

| | v.test | Mean in category | Overall mean | sd in category | Overall sd | p.value |
|---|---|---|---|---|---|---|
| Abstention | 79.165836 | 37.935012778 | 25.18359401 | 13.50818761 | 9.3131336 | 0.000000e+00 |
| Melenchon | 62.545344 | 25.232931727 | 16.61182604 | 10.65358458 | 7.9697208 | 0.000000e+00 |
| Arthaud | -2.908255 | 0.009857612 | 0.01643239 | 0.09879494 | 0.1307145 | 3.634519e-03 |
| Poutou | -8.381755 | 0.025556773 | 0.06183594 | 0.15780882 | 0.2502635 | 5.214414e-17 |
| Roussel | -22.334009 | 0.740051114 | 1.11667657 | 0.84827777 | 0.9750303 | 1.727179e-110 |
| Hidalgo | -32.246364 | 0.364001460 | 0.77153039 | 0.52609542 | 0.7307231 | 3.956028e-228 |
| LePen | -36.516704 | 9.804308142 | 14.65887943 | 6.20429573 | 7.6866027 | 6.023630e-292 |
| Lasalle | -44.525563 | 0.357794816 | 1.39992081 | 0.54090140 | 1.3532746 | 0.000000e+00 |
| Dupont-Aignant | -44.625832 | 0.320920044 | 0.88431334 | 0.48973072 | 0.7299624 | 0.000000e+00 |
| Jadot | -45.688627 | 1.621394670 | 3.20359005 | 1.58377893 | 2.0022907 | 0.000000e+00 |
| Zemmour | -50.795464 | 2.592917123 | 4.82452320 | 1.58938412 | 2.5401989 | 0.000000e+00 |
| Pecresse | -51.835441 | 1.161737861 | 3.04711938 | 0.90420100 | 2.1030402 | 0.000000e+00 |
| Macron | -66.379799 | 12.245345016 | 20.42836402 | 4.32636821 | 7.1277549 | 0.000000e+00 |

$`2`

| | v.test | Mean in category | Overall mean | sd in category | Overall sd | p.value |
|---|---|---|---|---|---|---|
| LePen | 91.710953 | 20.39600688 | 14.65887943 | 5.2810845 | 7.6866027 | 0.000000e+00 |
| Dupont-Aignant | 60.826998 | 1.24566971 | 0.88431334 | 0.6846908 | 0.7299624 | 0.000000e+00 |
| Lasalle | 59.984385 | 2.06055798 | 1.39992081 | 1.4974201 | 1.3532746 | 0.000000e+00 |
| Roussel | 33.151965 | 1.37974349 | 1.11667657 | 1.1114883 | 0.9750303 | 5.305853e-241 |
| Zemmour | 13.349009 | 5.10048922 | 4.82452320 | 2.1492514 | 2.5401989 | 1.200345e-40 |
| Poutou | 13.005467 | 0.08832474 | 0.06183594 | 0.2987477 | 0.2502635 | 1.139012e-38 |
| Arthaud | 11.027419 | 0.02816343 | 0.01643239 | 0.1709426 | 0.1307145 | 2.818385e-28 |
| Hidalgo | -6.672023 | 0.73185244 | 0.77153039 | 0.6777666 | 0.7307231 | 2.523003e-11 |
| Pecresse | -20.729987 | 2.69231786 | 3.04711938 | 1.2249890 | 2.1030402 | 1.858387e-95 |
| Abstention | -21.077970 | 23.58601084 | 25.18359401 | 4.7456641 | 9.3131336 | 1.267022e-98 |
| Macron | -22.371344 | 19.13063599 | 20.42836402 | 4.4625009 | 7.1277549 | 7.484785e-111 |
| Jadot | -42.726790 | 2.50733836 | 3.20359005 | 1.1780723 | 2.0022907 | 0.000000e+00 |
| Melenchon | -55.469763 | 13.01401560 | 16.61182604 | 3.7232099 | 7.9697208 | 0.000000e+00 |

$`3`

| | v.test | Mean in category | Overall mean | sd in category | Overall sd | p.value |
|---|---|---|---|---|---|---|
| Jadot | 83.478537 | 5.182436611 | 3.20359005 | 1.69064651 | 2.0022907 | 0.000000e+00 |
| Macron | 78.729819 | 27.071943929 | 20.42836402 | 5.70193274 | 7.1277549 | 0.000000e+00 |
| Pecresse | 64.973776 | 4.664811379 | 3.04711938 | 2.48877021 | 2.1030402 | 0.000000e+00 |
| Hidalgo | 33.749153 | 1.063492063 | 0.77153039 | 0.78357237 | 0.7307231 | 1.100108e-249 |
| Zemmour | 27.591951 | 5.654298083 | 4.82452320 | 2.81224675 | 2.5401989 | 1.389874e-167 |
| Melenchon | 7.858864 | 17.353329210 | 16.61182604 | 7.07200639 | 7.9697208 | 3.876341e-15 |
| Poutou | -7.024847 | 0.041022470 | 0.06183594 | 0.20041008 | 0.2502635 | 2.143005e-12 |
| Arthaud | -9.419706 | 0.001855288 | 0.01643239 | 0.04303307 | 0.1307145 | 4.523540e-21 |
| Roussel | -17.108230 | 0.919191919 | 1.11667657 | 0.64053987 | 0.9750303 | 1.288567e-65 |
| Lasalle | -27.561022 | 0.958359101 | 1.39992081 | 0.75560007 | 1.3532746 | 3.264966e-167 |
| Dupont-Aignant | -28.381381 | 0.639043496 | 0.88431334 | 0.59596755 | 0.7299624 | 3.432919e-177 |
| Abstention | -42.709802 | 20.474541332 | 25.18359401 | 3.85090099 | 9.3131336 | 0.000000e+00 |
| LePen | -68.169931 | 8.455370027 | 14.65887943 | 4.02955193 | 7.6866027 | 0.000000e+00 |

FIGURE 9 – Caractéristiques des différents groupes

Dans cette liste, nous retrouvons les différentes caractéristiques significatives décrivant les groupes. Ce qui est intéressant pour nous, ce sont les 3 premières colonnes : **v.test**, **Mean in category** et **Overall mean**.

La colonne **Mean in category** indique la moyenne de la variable pour les individus du groupe. Overall mean**Overall mean**ne de la variable sur l'ensemble des individus. Enfin, la variable **v.test** décrit la significativité de la variable au sein du groupe. Une variable très fortement significative dans un groupe indique que les individus du groupe prennent une valeur significativement forte sur cette variable comparé à l'ensemble des individus. De la même manière, une variable très négativement significative dans un groupe indique que les individus du groupe prennent une valeur significativement faible sur cette variable comparé à l'ensemble des individus.

Ainsi, nous pouvons observer que les villes du **groupe 1** ont un fort taux d'abstention et Macron y fait de mauvais score. Cependant, les villes du **groupe 1** sont celles ou Jean-Luc Mélenchon y fait ses meilleurs scores. Enfin, Jadot fait de gros scores dans les villes ou le taux d'abstention

est faible. Ce sont également les mêmes villes où Le Pen fait ses plus mauvais score.

Pour savoir quelles villes contribuent le plus à chaque axe, nous utilisons l'instruction :

```
vote.clust$data.clust
```

Nous obtenons :

```
$`1`
                                    Cla/Mod    Mod/Cla     Global       p.value    v.test
Libellé.de.la.commune=Tourcoing    91.836735 1.64293538 0.32336831 2.760268e-29 11.234508
Libellé.de.la.commune=Marseille    45.051195 4.81927711 1.93361051 7.233776e-27 10.731594
Libellé.de.la.commune=Saint-Denis  97.222222 1.27783863 0.23757672 2.491940e-25 10.399517
Libellé.de.la.commune=Argenteuil   96.969697 1.16830960 0.21777866 3.988531e-23  9.904275
Libellé.de.la.commune=Roubaix      96.551724 1.02227090 0.19138124 3.407415e-20  9.205293
Libellé.de.la.commune=Montreuil    80.000000 1.31434830 0.29697090 2.489681e-19  8.989230
Libellé.de.la.commune=Aubervilliers 100.000000 0.83972253 0.15178513 7.588873e-18  8.605645
Libellé.de.la.commune=Fort-de-France 100.000000 0.80321285 0.14518577 4.226251e-17  8.406445
Libellé.de.la.commune=Pantin       95.454545 0.76670318 0.14518577 4.331114e-15  7.844954
Libellé.de.la.commune=Lille        64.406780 1.38736765 0.38936184 4.984156e-15  7.827311


$`2`
                                       Cla/Mod    Mod/Cla     Global       p.value
Libellé.de.la.commune=La Ciotat    100.0000000 0.19833399 0.09899030  2.950592e-05
Libellé.de.la.commune=Castres       91.3043478 0.27766759 0.15178513  3.426398e-05
Libellé.de.la.commune=Agde         100.0000000 0.18511173 0.09239095  5.917209e-05
Libellé.de.la.commune=Draguignan   100.0000000 0.17188946 0.08579159  1.186577e-04
Libellé.de.la.commune=Narbonne      90.0000000 0.23800079 0.13198707  2.131141e-04
Libellé.de.la.commune=Mauguio      100.0000000 0.15866720 0.07919224  2.379282e-04
Libellé.de.la.commune=Gujan-Mestras 100.0000000 0.15866720 0.07919224  2.379282e-04
Libellé.de.la.commune=Aubagne      100.0000000 0.15866720 0.07919224  2.379282e-04
Libellé.de.la.commune=Six-Fours-les-Plages 100.0000000 0.14544493 0.07259289  4.770536e-04
Libellé.de.la.commune=Saint-Laurent-du-Var 100.0000000 0.14544493 0.07259289  4.770536e-04


$`3`
                                    Cla/Mod     Mod/Cla     Global       p.value    v.test
Libellé.de.la.commune=Paris        84.606742 15.52257267 5.87342440 3.903561e-247 33.574877
Libellé.de.la.commune=Lyon         98.400000  2.53555968 0.82491916  1.826889e-58 16.120646
Libellé.de.la.commune=Bordeaux     83.050847  2.02020202 0.77872369  2.108728e-30 11.459440
Libellé.de.la.commune=Rennes       87.500000  1.73160173 0.63353791  1.269019e-29 11.302937
Libellé.de.la.commune=Toulouse     75.781250  1.99958771 0.84471722  2.703270e-24 10.169849
Libellé.de.la.commune=Boulogne-Billancourt 97.959184  0.98948670 0.32336831  5.238373e-23  9.876988
Libellé.de.la.commune=Nantes       90.322581  1.15440115 0.40915990  1.094761e-21  9.567543
Libellé.de.la.commune=Saint-Maur-des-Fossés 97.500000  0.80395795 0.26397413  1.283340e-18  8.807180
Libellé.de.la.commune=Versailles   94.871795  0.76272933 0.25737478  1.638739e-16  8.245935
Libellé.de.la.commune=Levallois-Perret 100.000000  0.49474335 0.15838448  1.291661e-12  7.095198
```

FIGURE 10 – Tableau représentant les 10 villes qui contribuent le plus pour chaque catégorie

Remarque : Nous avons pris une partie des villes qui contribuent le plus, soit les 10 villes qui contribuent le plus.

Pour répondre à la question, les bureaux de votes qui contribuent le plus à la **1e catégorie** sont Tourcoing, Marseille, Saint-Denis ; pour la **2e catégorie** La Ciotat, Castres, Agde et pour la **3e catégorie** Paris, Lyon, Bordeaux.

En analysant la Figure 8, nous pouvons donc décrire les bureaux de votes d'un point de vue géographique.

Nous avons représenté sur la carte de la France métropolitaine les partis politiques de chaque bureaux de votes (de façon très très synthétique) :

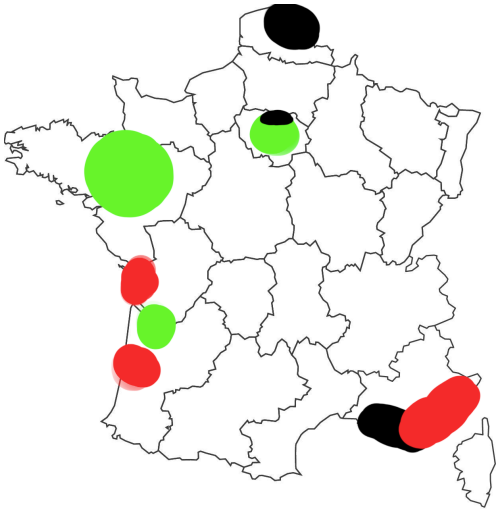


FIGURE 11 – Carte France métropolitaine du 1e tour des élections présidentielles de 2022 par catégorie : en rouge les villes qui votent le plus à droite, en noir les villes qui votent le plus à l'extrême gauche (et qui s'absente) et en vert les villes qui votent à gauche

## 1.6 Question 6

**Déduire de la question précédente une interprétation des axes principaux retenus.**

Ainsi, le premier axe représente le parti politique de gauche (et l'abstention) et le deuxième axe représente le parti politique de droite.

Nous pouvons de plus voir que les villes populaires **(classe 1)** votent principalement pour des candidats d'extrême gauche..

## 1.7 Question 7

**Y-a-t'il des bureaux de votes mal représentés dans la projection ? Que peut-on dire de ces bureaux ?**

Pour savoir quels bureaux de votes sont mals représentés dans la projection, nous utilisons de nouveau l'instruction :

```
vote.clust$data.clust
```

Nous regardons cette fois-ci les villes qui contribuent le moins, soient :

```
$`1`
Libellé.de.la.commune=Marcq-en-Baroeul          0.000000 0.00000000 0.13198707 1.849523e-02 -2.355552
Libellé.de.la.commune=Gap                       0.000000 0.00000000 0.13198707 1.849523e-02 -2.355552
Libellé.de.la.commune=Biarritz                  0.000000 0.00000000 0.13198707 1.849523e-02 -2.355552
Libellé.de.la.commune=Aix-les-Bains             0.000000 0.00000000 0.13198707 1.849523e-02 -2.355552
Libellé.de.la.commune=Levallois-Perret          0.000000 0.00000000 0.15838448 8.320846e-03 -2.638766
Libellé.de.la.commune=Versailles                2.564103 0.03650968 0.25737478 4.415739e-03 -2.846827
Libellé.de.la.commune=Saint-Maur-des-Fossés     2.500000 0.03650968 0.26397413 3.691016e-03 -2.903428
Libellé.de.la.commune=Les Sables-d'Olonne       0.000000 0.00000000 0.21117930 1.682978e-03 -3.141129
Libellé.de.la.commune=Boulogne-Billancourt      2.040816 0.03650968 0.32336831 7.224382e-04 -3.380919
Libellé.de.la.commune=Lyon                      1.600000 0.07301935 0.82491916 5.932253e-09 -5.818656

$`2`
Libellé.de.la.commune=Angers                    1.7543860 0.01322227 0.37616314  4.093300e-16  -8.135769
Libellé.de.la.commune=Lille                     1.6949153 0.01322227 0.38936184  1.054018e-16  -8.298536
Libellé.de.la.commune=Villeurbanne              1.4925373 0.01322227 0.44215667  4.575086e-19  -8.922105
Libellé.de.la.commune=Nantes                    0.0000000 0.00000000 0.40915990  2.137229e-19  -9.005993
Libellé.de.la.commune=Toulouse                  8.5937500 0.14544493 0.84471722  7.178373e-24 -10.074281
Libellé.de.la.commune=Montpellier               3.9215686 0.05288907 0.67313403  8.032844e-25 -10.287385
Libellé.de.la.commune=Rennes                    0.0000000 0.00000000 0.63353791  1.107361e-29 -11.314894
Libellé.de.la.commune=Bordeaux                  0.8474576 0.01322227 0.77872369  2.853021e-34 -12.206975
Libellé.de.la.commune=Lyon                      0.0000000 0.00000000 0.82491916  1.756833e-38 -12.972298
Libellé.de.la.commune=Paris                     0.0000000 0.00000000 5.87342440 5.781162e-280 -35.753769

$`3`
Libellé.de.la.commune=Roubaix                  3.448276  0.02061431 0.19138124  2.140303e-04 -3.701853
Libellé.de.la.commune=Fort-de-France           0.000000  0.00000000 0.14518577  2.042599e-04 -3.713689
Libellé.de.la.commune=Calais                   0.000000  0.00000000 0.14518577  2.042599e-04 -3.713689
Libellé.de.la.commune=Castres                  0.000000  0.00000000 0.15178513  1.387742e-04 -3.810343
Libellé.de.la.commune=Aubervilliers            0.000000  0.00000000 0.15178513  1.387742e-04 -3.810343
Libellé.de.la.commune=Argenteuil               3.030303  0.02061431 0.21777866  5.105293e-05 -4.050754
Libellé.de.la.commune=Marseille               20.477816  1.23685838 1.93361051  9.553302e-06 -4.427045
Libellé.de.la.commune=Saint-Denis              0.000000  0.00000000 0.23757672  9.092985e-07 -4.910316
Libellé.de.la.commune=Nouméa                   3.773585  0.04122861 0.34976572  4.538749e-07 -5.044851
Libellé.de.la.commune=Tourcoing                2.040816  0.02061431 0.32336831  1.492399e-07 -5.253495
```

FIGURE 12 – Tableau représentant les 10 villes qui contribuent le moins pour chaque catégorie

Remarque : Nous avons pris une partie des villes qui contribuent le moins, soit les 10 villes qui contribuent le moins.

Les bureaux de votes qui contribuent le moins à la **1e catégorie** sont Lyon, Boulogne-Billancourt, Les Sables-d'Orlonne ; **2e catégorie** Paris, Lyon, Bordeaux ; pour la **3e catégorie** Tourcoing, Nouméa, Saint-Denis.

Nous remarquons que les bureaux de votes qui contribuent le moins à la 2e catégorie sont ceux qui contribuent le plus à la 3e catégorie et les bureaux de votes qui contribuent le moins à la 3e catégorie sont ceux qui contribuent le plus à la 1e catégorie.

Ainsi, lorsque la 1e catégorie de villes votent pour un candidat, celui-ci ne sera que très peu voté par les villes de la 3e catégorie et lorsque la 3e catégorie de villes votent pour un candidat, celui-ci ne sera que très peu voté par les villes de la 2e catégorie.

Nous pouvons en conclure que les bureaux de votes qui votent majoritairement pour l'extrême gauche ne votent que très peu pour la droite et les bureaux de votes qui votent pour l'extrême droite ne votent que très peu pour l'extrême gauche (ce qui nous ramène au résultat de la Figure 9).

# 2 Exercice 2 (Régression prix d'un bien immobilier)

Pour finir, nous cherchons dans ce dernier exercice à déterminer le prix d'un bien immobilier à l'aide de divers descripteurs (taille, nombre de pièces, jardin, localisation, etc.).

Nous nous plaçons pour cela dans la peau d'un statisticien immobilier. Nous avons une base de données trouvables sur le site du gouvernement qui est l'ensemble des transactions faites en France et nous avons envie d'avoir une idée du nombre de pièces, de la taille, etc.
Le but est ainsi de faire une estimation d'un prix raisonnable.

## 2.1 Question 1

Supposons qu'il existe $f$ telle que :

$$\text{Prix de vente} = f(\text{descripteurs}) + \varepsilon \tag{1}$$

où $\varepsilon$ est une variable indépendante des descripteurs. Nous avons que l'hypothèse sur $f$ est toujours vérifiée dans une régression.

Nous voulons ici **proposer une solution pour régler ce problème**. Nous nous **interrogerons à cet égard sur notre perception de la variance de $\varepsilon$.**

Nous utilisons pour cela le **modèle régression linaire** que nous réaliserons sur R. A l'aide de la commande `summary()`, nous obtenons la sortie suivante qui pourrait illustrer la régression linéaire :

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -2.902e+06  2.463e+05 -11.786  < 2e-16 ***
numero_disposition       9.287e+05  1.218e+05   7.626 2.43e-14 ***
nombre_lots             -1.009e+06  3.948e+05  -2.557   0.0106 *
code_type_local          1.886e+06  3.679e+04  51.266  < 2e-16 ***
nombre_pieces_principales  1.202e+05  1.625e+04   7.397 1.40e-13 ***
surface_terrain          9.532e+01  3.067e+00  31.085  < 2e-16 ***
longitude               -1.181e+05  3.670e+03 -32.180  < 2e-16 ***
latitude                 1.908e+04  3.985e+03   4.788 1.68e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

FIGURE 13 – Résultats du modèle de régression linéaire

Tout d'abord, étudions la p-valeur ($Pr > |t|$) d'après le tableau de sortie ci-dessus (Figure 13). Les facteurs dont la p-valeur sont inférieur à 0.001 sont très significatif ($p < 0.001$) , les facteurs dont la p-valeur sont inférieur à 0.01 sont significatif ($p < 0.01$) , p-valeur inférieur à 0.1 alors tendance à la signification ($p < 0.1$) et enfin p-valeur supérieur ou égale à 0.1 alors non significatif ($p \geq 0.1$).
Les facteurs dont la p-valeur est inférieure ou égale à 0.05 sont généralement considérés comme significatifs dans un modèle. Nous utilisons donc un seuil de significativité de 0.05.

Ainsi, avec un seuil de significativité de 0.05, nous pouvons conclure que les données de la Figure 13 peuvent utiliser la régression linéaire suivante :

$$\begin{aligned} \text{valeur\_fonciere} = & -2902000 \\ & + \text{numero\_disposition} \times 928700 \\ & + \text{nombre\_lots} \times -1009000 \\ & + \text{code\_type\_local} \times 1886000 \\ & + \text{surface\_terrain} \times 95320 \\ & + \text{longitude} \times -118100 \\ & + \text{latitude} \times 19080 \end{aligned} \tag{2}$$

Analysons maintenant **la variance de** $\varepsilon$.

```
Residuals:
       Min         1Q     Median         3Q        Max
-352812523   -1441097   -1019209    -677626  721439573
```

FIGURE 14 – Résultats de régression linéaire

Nous pouvons voir d'après la Figure 14 que la moyenne des résidus est très proche de 0, et que la médiane des prix des résidus est de **-1019209**, très loin de 0. Les chiffres tels que les quantiles (25% et 75%) sont très différents des autres chiffres. Par conséquent, nous pouvons en conclure que le modèle de surface est très inégal.

Dans la pratique, si tous les facteurs qui **influencent le prix de l'immobilier** ne suivent pas une trajectoire idéale, cela signifie que la part due à l'erreur est importante. Par conséquent, dans le groupe où le prix de l'immobilier est évalué, si l'erreur n'est pas bien contrôlée, cela montrera que le modèle de prédiction n'est pas précis.

Des **solutions** peuvent être apportés pour résoudre ce problème, comme par exemple éliminer les données qui ne conviennent pas, s'assurer que toute relation entre les variables explicatives et la variable dépendante est linéaire (si ce n'est pas le cas, le modèle de régression linéaire peut ne pas être approprié), s'assurer que les données sont propres et traitées avant de ne pas inclure de données incorrectes qui pourraient affecter l'efficacité du modèle. Parfois, la régression linéaire simple n'est pas suffisante pour capturer des relations complexes entre les variables. Dans ces cas, nous pouvons considérer des modèles plus complexes par rapport à la régression linéaire.

## 2.2 Question 2

Nous voulons **proposer un modèle d'estimation du prix de vente d'un bien immobilier**.

Nous utilisons dans cette question le **modèle Lasso** qui est une des méthodes qui vient pallier les manques (instabilité de l'estimation et manque de fiabilité de la prévision) de la régression linéaire dans un contexte de grande dimension. L'avantage principal de cette régression réside dans sa capacité à effectuer une sélection de variables, ce qui peut s'avérer précieux en présence d'un grand nombre de variables.

Voici les étapes que nous avons suivi pour réaliser la régression Lasso.

**Premièrement** nous divisions les données en un ensemble d'apprentissage et un ensemble de test selon un ratio de 80/20. La variable réponse `valeur_fonciere` et des variables prédictives telles que `"numero_disposition"`, `"nombre_lots"`, `"code_type_local"`, `"nombre_pieces_principales"` `"surface_terrain"`, `"longitude"`, `"latitude"` sont sélectionnées à partir de l'ensemble de données pour créer `X_train`, `X_test`, `Y_train`, `Y_test`.

**Deuxièmement**, nous faisons la normalisation des variables prédictives avec la moyenne et l'écart-type de l'ensemble d'apprentissage pour améliorer l'efficacité du modèle Lasso.

```
set.seed(123)  # For reproducible results
split <- sample.split(Y, SplitRatio = 0.8)

train_data <- subset(data, split == TRUE)
test_data <- subset(data, split == FALSE)

# Splitting predictors and response for training and test sets
X_train <- train_data[, c("numero_disposition" ,"nombre_lots" , "code_type_local" , "nombre_pieces_principales" , "surface_terra
Y_train <- train_data$valeur_fonciere

X_test <- test_data[, c("numero_disposition" ,"nombre_lots" , "code_type_local" , "nombre_pieces_principales" , "surface_terrain
Y_test <- test_data$valeur_fonciere
```

**Troisième étape**, nous utilisons la validation croisée pour trouver la valeur optimale de lambda pour le modèle. Ensuite, nous entraînons le modèle Lasso avec le lambda optimal et évaluons sa performance sur l'ensemble du test.

```
# Standardize the predictors for Lasso
X_train_matrix <- as.matrix(scale(X_train))

# Fit Lasso Model
cv.lasso <- cv.glmnet(X_train_matrix, Y_train, alpha = 1)
best.lambda <- cv.lasso$lambda.min

lasso.model <- glmnet(X_train_matrix, Y_train, alpha = 1, lambda = best.lambda)
```

**Au final**, nous calculons les indicateurs de performance tels que le MSE et le R-carré pour comparer le modèle Lasso avec un modèle de régression linéaire simple (voir analyse dans la question suivante).

```r
# Preparing test data for Lasso (standardize using training mean and sd)
X_test_matrix <- as.matrix(scale(X_test, center = attr(X_train_matrix, "scaled:center"), scale = attr(X_train_matrix, "scaled:sc

# Predictions
predictions_lasso <- predict(lasso.model, s = best.lambda, newx = X_test_matrix)
predictions_linear <- predict(linearmodel, newdata = test_data)

# Evaluate predictions, e.g., using RMSE, MAE, R²
# You'll need to write or find functions to calculate these metrics as per your requirement
```

```r
# MSE

library(Metrics)

mse_lasso <- mse(Y_test, predictions_lasso)
mse_linear <- mse(Y_test, predictions_linear)

# RMSE
rmse_lasso <- sqrt(mse_lasso)
rmse_linear <- sqrt(mse_linear)

cat("rmse_lasso = ", rmse_lasso)
cat("\nrmse_linear = ", rmse_linear)

# MAE
mae_lasso <- mae(Y_test, predictions_lasso)
mae_linear <- mae(Y_test, predictions_linear)

cat("\nmae_lasso = ", mae_lasso)
cat("\nmae_linear = ", mae_linear)

# R² for Linear Regression
r_squared_linear <- summary(linearmodel)$r.squared

# R² for Lasso
# Sum of Squares Total
SST <- sum((Y_test - mean(Y_test))^2)
# Sum of Squares Residual
SSR <- sum((Y_test - predictions_lasso)^2)
# R²
r_squared_lasso <- 1 - SSR/SST

cat("\nr_squared_linear = ", r_squared_linear)
cat("\nr_squared_lasso = ", r_squared_lasso)
```

```
rmse_lasso =  20537150
rmse_linear =  20536877
mae_lasso =  2997518
mae_linear =  3000223
r_squared_linear =  0.008791904
r_squared_lasso =  0.008961122
```

## 2.3 Question 3

Pour finir, nous voulons **donner une estimation de la performance du modèle choisi**. **Avec le modèle de régression linéaire :**

Pour la partie résiduelle nous trouvons que la moyenne des résidus devrait idéalement être nulle. Cependant, dans notre cas, elle est significativement éloignée de zéro, avec une moyenne d'environ **-1029923**. Cela montre un écart considérable entre les valeurs observées et les prédictions, suggérant que le modèle pourrait ne pas être adapté aux données. Une telle erreur importante peut indiquer que le modèle linéaire n'est pas approprié pour les données, possiblement en raison de relations non linéaires entre les variables que le modèle linéaire ne peut pas capturer.

De plus, une erreur aussi grande pourrait également signaler la présence de points de données anormalement différents par rapport au reste des données. L'échelle de l'erreur doit être considérée dans le contexte de l'échelle de la variable dépendante. Si la variable dépendante varie dans une large gamme (par exemple, des millions ou des milliards), alors une grande erreur peut être moins préoccupante.

```
rmse_lasso =  20537150
rmse_linear =  20536877
mae_lasso =  2997518
mae_linear =  3000223
r_squared_linear =  0.008791904
r_squared_lasso =  0.008961122
```

FIGURE 15 – Aperçu des résultats du modèle Lasso et du modèle de régression linéaire

Nous considérons que le $R^2$ est de **0.008791904**, indiquant que le modèle explique seulement environ 0.88% de la variance du marché immobilier, une valeur très faible.
Les valeurs de $R^2$ et $R^2$ ajusté fournissent une indication de l'adéquation du modèle aux données, mais elles ne sont pas les seuls indicateurs pertinents.
Il est important d'examiner la valeur P des prédicteurs dans le tableau de sortie. Les facteurs avec des valeurs P faibles sont considérés comme significatifs.
L'analyse de la partie résiduelle est cruciale pour confirmer que les résidus sont distribués de manière aléatoire. Des motifs non aléatoires peuvent indiquer un mauvais ajustement du modèle ou l'omission de variables prédictives importantes.

**Avec le modèle de régression Lasso :**
Les résultats indiquent que le modèle Lasso a une performance légèrement supérieure par rapport au modèle de régression linéaire multiple, selon les valeurs de RMSE et MAE. Le coefficient de détermination $R^2$ est également légèrement plus élevé pour le modèle Lasso, ce qui suggère une meilleure adéquation du modèle aux données de test par rapport à la régression linéaire multiple. Cependant, les différences entre les deux modèles sont minimes, ce qui implique que leurs performances sont comparables sur cet ensemble de données.
Pour une interprétation plus poussée, il serait nécessaire d'examiner l'ampleur des erreurs par rapport à la variabilité des données et de tenir compte du contexte spécifique des données analysées.
La valeur de l'erreur absolue moyenne (MAE) du Lasso est meilleure que celle du modèle linéaire, soit $2997518 < 3000223$.
La valeur du coefficient de détermination $R^2$ du Lasso est également meilleure que celle du modèle linéaire, soit $0.008961122 > 0.008791904$.

**Conclusion sur le modèle :** Bien que le modèle actuel ne soit pas parfait, il est possible d'améliorer sa précision en effectuant des ajustements supplémentaires, comme l'inclusion de termes d'interaction ou l'exploration de relations non linéaires.

# Conclusion

Pour l'Exercice 1, l'analyse des profils des bureaux de vote à travers une analyse de données nous a permis de discerner des tendances intéressantes. Les résultats obtenus illustrent des liens géographiques et des profils politiques distincts, reflétant une hétérogénéité marquée au sein des différentes régions de la France. L'application de méthodes statistiques appropriées, telles que l'ACP, a mis en lumière les variables les plus significatives, nous permettant ainsi d'identifier les bureaux de vote les plus contributifs à chaque axe de l'analyse. En outre, l'interprétation des axes principaux a donné un aperçu des caractéristiques politiques prédominantes, révélant les nuances dans le paysage politique français.

Pour l'Exercice 2, l'estimation des prix des biens immobiliers à l'aide de modèles de régression montre que le modèle Lasso, bien qu'offrant une amélioration marginale par rapport à la régression linéaire multiple, est légèrement plus performant selon les métriques d'erreur moyenne absolue et du coefficient de détermination. Cette analyse démontre l'importance de sélectionner un modèle adapté pour prédire les prix immobiliers en prenant en compte la variance des descripteurs et en proposant des solutions pour atténuer l'impact de cette dernière sur la régression.

En somme, ces deux exercices mettent en évidence la puissance des outils statistiques et des modèles de régression dans l'analyse de données complexes, que ce soit pour comprendre le comportement électoral ou pour estimer la valeur des biens immobiliers. Ils soulignent également la nécessité d'une interprétation minutieuse des résultats pour tirer des conclusions éclairées.