Rapport de projet :

Multicolinéarité et régression PLS

Membres du groupe :

Cheddadi Radja
Abichou Nour Elhouda

2022/2023

Table de matière :

Introduction

L'analyse des élections et des résultats électoraux est une tâche complexe et importante pour les partis politiques, les médias et les électeurs. Dans ce rapport, nous présentons une analyse approfondie des données et des modèles utilisés pour prédire l'issue d'une élection entre différents candidats.

Nous commencerons par une analyse exploratoire des données, en utilisant des techniques d'analyse univariée et bivariée. Ensuite, nous utiliserons l'analyse de composantes principales pour identifier les dimensions les plus importantes des données.

Nous présenterons ensuite plusieurs modèles unidimensionnels, notamment la régression linéaire, la régression de la composante principale et la régression partielle des moindres carrés.

Nous comparerons ensuite ces modèles et présenterons un modèle multidimensionnel basé sur la régression partielle des moindres carrés.

Enfin, nous conclurons en présentant les résultats de nos analyses et en discutant de leur pertinence pour prédire l'issue des élections.

1. Analyse exploratoire des données

Nous allons examiner les distributions univariées et bivariées de nos variables, afin de construire des modèles sur des bases raisonnables.

1.1 Analyse univariée

Nous avons mis en place des histogrammes et des box plots afin de nous aider à effectuer une analyse des points atypiques et des fréquences. On peut voir que les proportions de votes pour ROUSSEL varient considérablement d'un département à l'autre, allant de moins de 1 % à plus de 3 %. Le département où ROUSSEL a obtenu la proportion la plus élevée de voix est la Corrèze avec environ 3,4 % des votes, tandis que les départements où il a obtenu les proportions les plus faibles sont les départements d'Outre-Mer (Guadeloupe, Martinique, Guyane, La Réunion, Mayotte) ainsi que les départements de l'Est de la France (Bas-Rhin, Haut-Rhin). En moyenne, la proportion de votes pour ROUSSEL est d'environ 1,9 % dans l'ensemble des départements.

La fréquence la plus élevée de voix pour Macron a été enregistrée dans le département de Hauts-de-Seine, où il a remporté 28,81% des voix. La fréquence la plus faible de voix pour Macron a été enregistrée en Guadeloupe, où il n'a remporté que 5,74% des voix. Les départements où Macron a obtenu les deuxièmes et troisièmes pourcentages de voix les plus élevés sont Paris (27,25%) et Yvelines (25,60%), respectivement. Les départements où Macron a obtenu les deuxièmes et troisièmes pourcentages de voix les plus faibles sont Seine-Saint-Denis (13,87%) et Ariège (15,18%), respectivement. En général, Macron a obtenu une part importante de voix dans les départements de l'ouest de la France, tels que la Vendée, la Mayenne, la Sarthe et la Loire-Atlantique, ainsi que dans les départements de l'Île-de-France, tels que Paris, les Yvelines et l'Essonne. Il remporte également une part importante de voix dans les départements du nord-est de la France, tels que la Moselle et le Bas-Rhin.

Le candidat Lassalle a obtenu des scores très variables selon les départements, allant de 0,002% à Wallis et Futuna à 9,27% dans les Pyrénées-Atlantiques. En moyenne, sa fréquence de vote s'élève à environ 0,03%, mais cela cache des écarts importants entre les départements où il a bénéficié d'un soutien plus important. Les régions où Lassalle a obtenu les meilleurs scores se trouvent dans le sud-ouest de la France, notamment en Ariège (6,33%), Aveyron (6,76%), Lot (5,77%), Hautes-Pyrénées (8,32%) et Pyrénées-Atlantiques (9,27%). En revanche, dans certains départements, tels que la Guadeloupe et la Martinique, Lassalle n'a reçu qu'un faible soutien (0,32% et 0,38%, respectivement).

On peut observer que la proportion de votes en faveur de Mme Le Pen varie fortement d'un département à l'autre, allant de 4,27% à Paris à 27,96% dans le département de l'Aisne. On observe également que les départements du Nord et du Pas-de-Calais, ainsi que certains départements du sud de la France, ont donné un pourcentage élevé de votes à Mme Le Pen. En revanche, dans les départements de l'Ouest et du Sud-Ouest de la France, les pourcentages sont généralement plus faibles.

On peut observer que les fréquences d'observations du candidat Zemmour varient assez largement d'un département à l'autre, allant d'environ 3% à plus de 10% dans certains départements. On peut remarquer que la fréquence d'observations est assez élevée dans les Alpes-Maritimes, le Var et le Vaucluse, ainsi que dans une partie de l'Occitanie. Ces départements sont situés dans le sud-est de la France et ont souvent été considérés comme des bastions de l'extrême droite.
D'autres départements avec une fréquence d'observations élevée comprennent l'Ain, les Bouches-du-Rhône, le Rhône et Paris. On peut noter que la fréquence d'observations dans la plupart des départements est inférieure à 5%, ce qui suggère que la popularité de Zemmour en tant que candidat est encore assez limitée dans l'ensemble de la France.

On peut observer que la fréquence des observations de Jean-Luc Mélenchon est variable selon les départements. En effet, elle varie de 0,0389% dans le département de Wallis-et-Futuna à 0,3358% en Seine-Saint-Denis, en passant par exemple par 0,2319% à Paris ou 0,2008% dans l'Essonne.

On peut toutefois noter que la fréquence des observations est généralement plus élevée dans les départements d'outre-mer et dans les départements les plus peuplés de la région parisienne. A l'inverse, elle est généralement plus faible dans les départements ruraux.

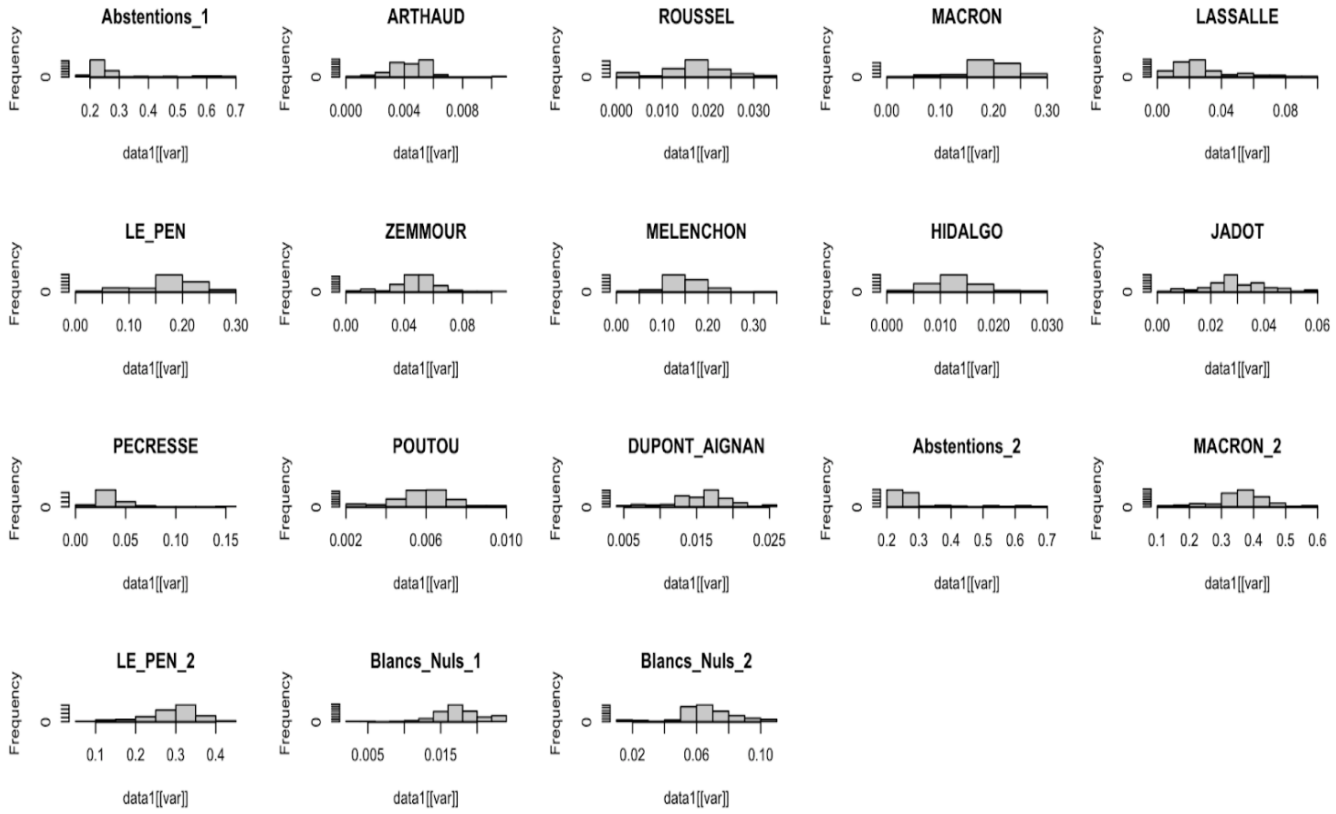
On constate pour Hidalo que les valeurs vont de 0,003% à Mayotte à 0,027% dans le département des Landes. La majorité des départements ont un pourcentage compris entre 0,01% et 0,02%.

Pour le candidat Jadot, onremarque que les fréquences varient considérablement d'un endroit à l'autre. La fréquence la plus élevée est observée à Paris avec 5,9%, suivie de la Haute-Savoie (5,3%), de l'Ille-et-Vilaine (5,5%), et de la Loire-Atlantique (5,7%). En général, les régions du nord-ouest de la France ont tendance à avoir des fréquences plus élevées pour Jadot, tandis que les régions du nord-est ont tendance à avoir des fréquences plus faibles. Les départements du sud-est de la France ont également tendance à avoir des fréquences plus faibles.
La fréquence des observations du candidat Pecresse varie selon les départements. La fréquence la plus élevée se trouve dans le département des Yvelines avec 6,37%, tandis que la fréquence la plus faible se trouve en Guyane avec 0,96%.

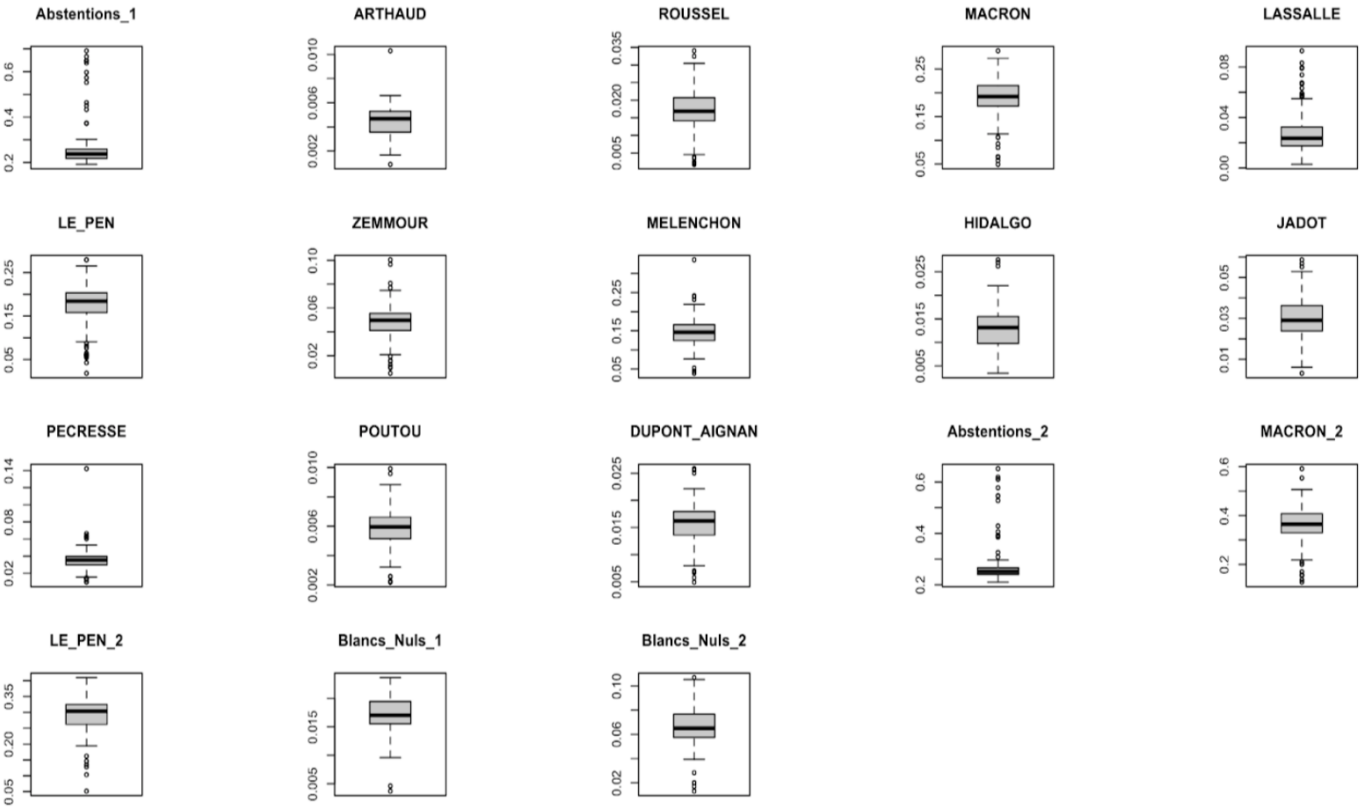
La fréquence des observations de POUTOU varie considérablement selon les départements. Le département avec la fréquence la plus élevée est Pyrénées-Atlantiques avec 9,58%, suivi de Saint-Pierre-et-Miquelon avec 9,91% et Côtes-d'Armor avec 8,26%. Les départements avec les fréquences les plus faibles sont Hauts-de-Seine avec 3,71%, Paris avec 4,19% et Alpes-Maritimes avec 3,84%. En général, les fréquences des observations de POUTOU se situent entre 4% et 8%, avec une moyenne d'environ 6%.
On peut observer que la fréquence des observations du candidat DUPONT_AIGNAN est assez variable entre les départements, allant de 0,0057 % à 0,0258 %. La moyenne de sa fréquence d'observation est d'environ 1,65 %. Les départements où il a obtenu les pourcentages les plus élevés sont le Haut-Rhin (2,58 %), la Haute-Savoie (2,51 %), le Jura (2,17 %) et les Vosges (2,21 %). En revanche, les départements où il a obtenu les pourcentages les plus faibles sont la Guyane (0,70 %), la Guadeloupe (0,67 %) et Mayotte (0,57 %).
Pour la fréquence des observations du candidat de Macron au deuxième tour. On remarque qu'il a obtenu un score relativement homogène sur l'ensemble du territoire avec une légère différence en fonction des départements. Dans le département de l'Ain, 31,1% des personnes ont voté pour Le Pen. Dans le département de l'Aisne, 40,97% des personnes ont voté pour Le Pen. Dans le département de Paris, seulement 10,35% des personnes ont voté pour Le Pen.

Globalement, les pourcentages les plus élevés sont concentrés dans l'est de la France, tandis que les pourcentages les plus faibles sont situés dans les départements d'outre-mer. Les observations dans le Guadeloupe, la Martinique, la Guyane, la Réunion et Mayotte peuvent être considérées comme des valeurs atypiques et extrêmes étant donné qu'elles sont significativement différentes de la majorité des données et de leur tendances.

Graphique 1 : histogrammes des variables



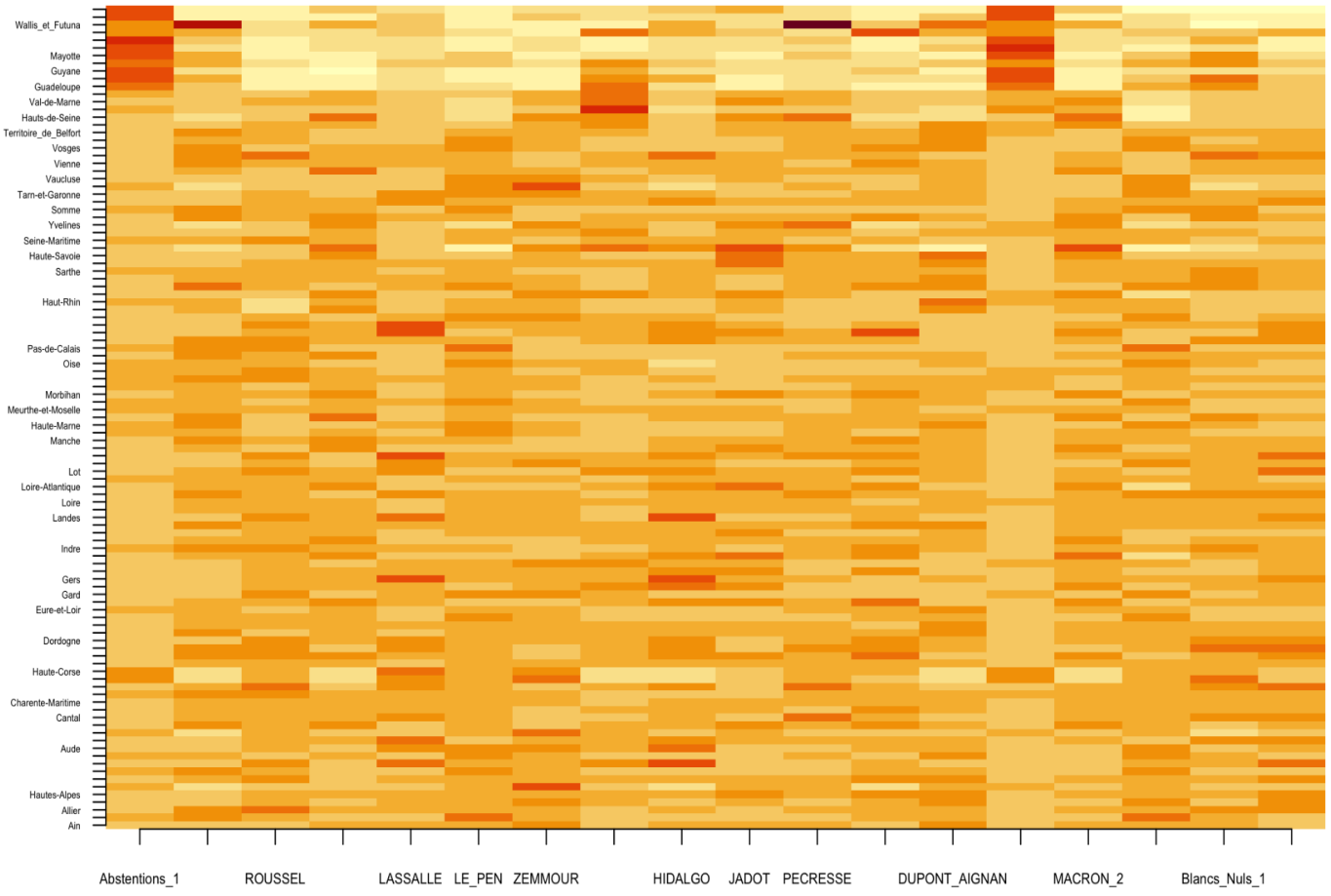
Graphique 2 : Boxplots des variables



Afin de confirmer notre constat, nous avons généré une visualisation sous forme de carte thermique des données standardisées. On observe dans la cartographie que les départements d'outre mer ont effectivement tendance à être différents des métropoles avec de très forts taux d'abstention et de très faibles taux de votes pour une grande majorité des candidats, contrairement au reste des départements qui affichent une tendance cohérente entre eux.

Afin que cette tendance contradictoire ne puisse pas interférer avec notre modélisation, nous allons choisir de nous concentrer que sur les départements en métropoles et de supprimer la Guadeloupe, la Martinique, la Guyane, la Réunion et Mayotte de notre analyse.

Graphique 3 : Carte thermique des variables



1.2 Analyse bivariée

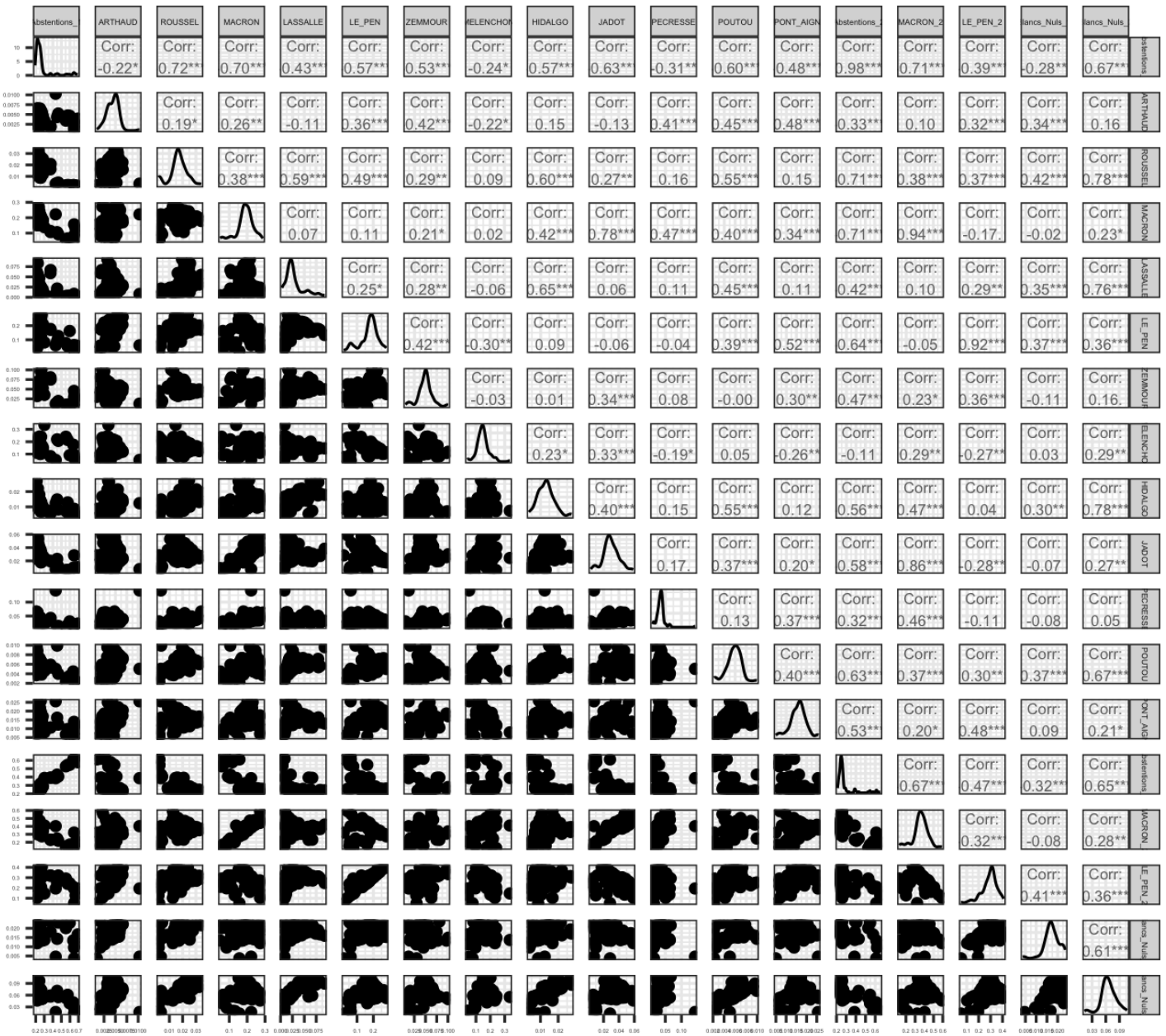
Le scatter plot que nous présentons à présent nous montre la relation entre les paires des 18 variables et va nous permettre de les évaluer visuellement. Le scatter plot est utile pour examiner la forme et la direction de la relation entre les deux variables, y compris les relations non linéaires et les valeurs aberrantes.

Nous pouvons constater l'existence de relations linéaires positives entre Roussel/Hidalgo, Macron/Jadot, Macron/Macron2, Lassal/Hadalgo, Macron2/Percease, BlancsNul2/Lassale, Abstentions1/Abstentions2, Lepen/lepen2, Macron2/Jadot…etc. Les points sur le graphique forment une ligne ascendante, cela signifie que lorsque la première variable augmente, la seconde variable augmente également.

Nous pouvons aussi constater l'existence de relations linéaires négatives entre Abstentions1/Macron et Abstentions1/Poutou. Les points sur le graphique forment une ligne descendante, ce qui indique une relation négative entre les deux variables. Cela signifie que lorsque la première variable augmente, la seconde variable diminue.

Le reste des paires de variables présentent des relation non linéaire ou les points sur le graphique ne suivent pas une ligne droite et prennent plusieurs formes, par exemple une courbe en U, une courbe en S, ou une forme irrégulière.

Graphique 4 : scatter plot des variables



Examinons à présent la matrice des corrélations linéaires simples. Notre matrice de corrélation montre les corrélations par paire entre les variables énumérées en haut et à gauche de la matrice.

Commençons par les variables expliquées. La corrélation entre Macron_2 et Macron est la corrélation positive la plus forte, avec un coefficient de 0,938, ce qui indique que le pourcentage de votes qui sont allés à Macron lors des deux tours de scrutin sont fortement corrélés. Suivi des corrélations avec avec Jadot, Hidalgo, Roussel et Poutou. La corrélation entre Le_Pen_2 et Le_Pen est la corrélation positive la plus forte avec un coefficient de 0,92. Suivi des corrélations de Le_Pen_2 avec avec les blanc_nuls_1, Dupont_Aignan et Zemmour.

La corrélation entre Blancs_Nuls_1 et Blancs_Nuls_2 est l'une des corrélations positives le plus fortes de la matrice, avec un coefficient de corrélation de 0,921. Cela suggère qu'il existe une relation très forte entre le nombre de bulletins blancs ou nuls aux deux tours de scrutin. On constate aussi une corrélation positive très forte entre Blancs_Nuls_2 et beaucoup de candidats du premier tour, entre autres Poutou, Roussel, Lassale et Hidalgo. Les corrélations entre Abstentions_2 et les autres variables sont presque toutes négatives, ce qui suggère que le pourcentage d'abstentions au second tour de scrutin est négativement corrélé avec le pourcentage de voix qui sont allées à ces candidats.

Hidalgo a des corrélations positives modérées avec Lassalle et Poutou, une corrélation positive faible avec Jadot, et une corrélation positive faible avec Pecresse. Nous pouvons voir qu'il existe une corrélation positives forte et entre Jadot et Macron (0,7792).

La corrélation négative la plus forte est entre Abstentions_1 et Roussel (-0.3302). Cela indique que plus le nombre d'abstentions au premier tour augmente, plus les résultats de vote pour Roussel diminuent. D'autres corrélations fortes incluent une corrélation positive entre Le Pen et Dupont_Aignan (0.5167) et entre Rousse et Hidalgo (0.5969), ainsi qu'une corrélation négative entre Zemmour et Arthaud (-0.4206). Ces corrélations suggèrent qu'il peut y avoir une certaine relation entre les résultats des votes pour ces candidats.

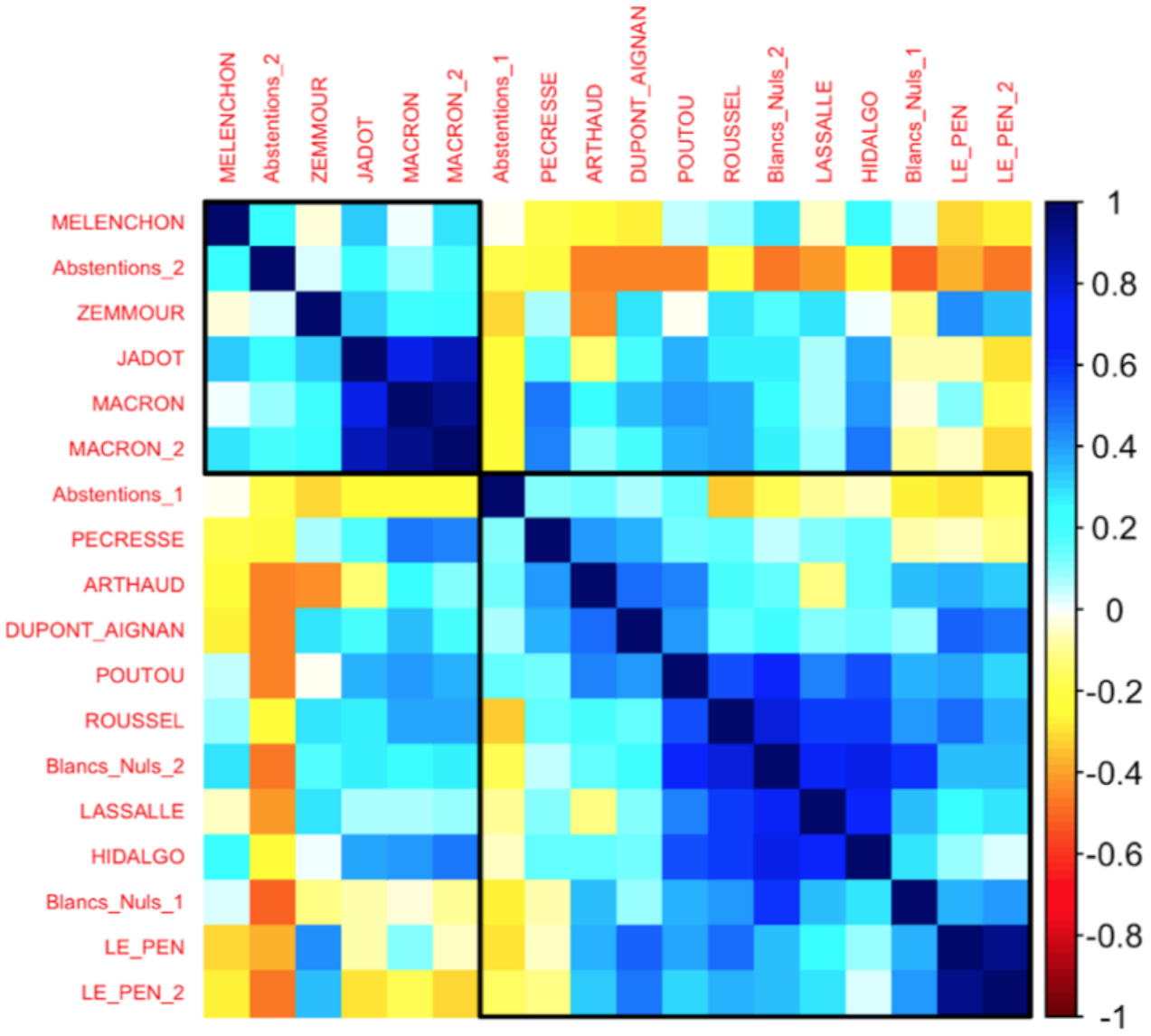
Jadot présente des corrélations positives faibles avec plusieurs autres variables, dont Hidalgo, Macron et Blancs_Nuls_1. Pecresse a une corrélation positive faible avec Dupont_Aignan et une corrélation positive très faible avec Jadot. Poutou a des corrélations positives modérées avec Lassalle et Hidalgo, et des corrélations positives faibles avec quelques autres variables.

La corrélation entre Hidalgo et ROUSSEL est la plus forte corrélation positive avec un coefficient de 0,597, ce qui indique qu'il peut y avoir des similitudes entre les électeurs qui ont soutenu ces deux candidats. La corrélation entre Poutou et Roussel est également forte, avec un coefficient de 0,552. Cela suggère qu'il peut y avoir un certain chevauchement dans les types d'électeurs qui ont soutenu ces deux candidats.

Les corrélations entre ZEMMOUR et Roussel, ZEMMOUR et LE_PEN_2, et ZEMMOUR et Dupont_Aignan sont toutes négatives, ce qui suggère que le pourcentage de voix qui sont allées à Zemmour est négativement corrélé avec le pourcentage de voix qui sont allées à ces autres candidats.

La corrélation entre Abstentions_1 et Roussel est négative et relativement forte, avec un coefficient de -0,33. Cela suggère que le pourcentage d'abstention au premier tour de scrutin peut être lié au pourcentage de voix qui sont allées à Roussel.

Graphique 5 : Matrice de corrélation



Il est intéressant de noter que les corrélations entre les candidats sont cohérentes avec leurs positions politiques respectives. Les candidats d'extrême-gauche, Nathalie Arthaud et Philippe Poutou, sont positivement corrélés, de même que les candidats de droite, Marine Le Pen et Nicolas Dupont-Aignan. Les candidats centristes, Jean Lassalle et Emmanuel Macron, sont également positivement corrélés.
Cela confirme l'idée selon laquelle les électeurs ont tendance à voter pour des candidats partageant leurs valeurs et leurs opinions politiques. En d'autres termes, les électeurs ont tendance à se regrouper en fonction de leurs idéologies politiques, ce qui peut contribuer à la polarisation de la société.

Les fortes corrélations entre les variables explicatives peuvent éventuellement poser un problème lors de la prédiction des variables Macron_2, Le_Pen_2, Abstentions_2 et Blancs_Nuls_2 en faisant des modèles de régressions linéaires multiples. Il convient de noter que la corrélation n'implique pas la causalité, et qu'une analyse plus approfondie serait nécessaire pour déterminer toute relation de cause à effet entre les variables.

2. Analyse de composantes principales

Dans le cadre de notre analyse multidimensionnelle nous allons adopter la méthode d'analyse de composantes principales, puisque nous avons constaté une potentielle multicolinéarité entre les variables de notre base de données.

2.2 Analyse multidimensionnelle de tour 1

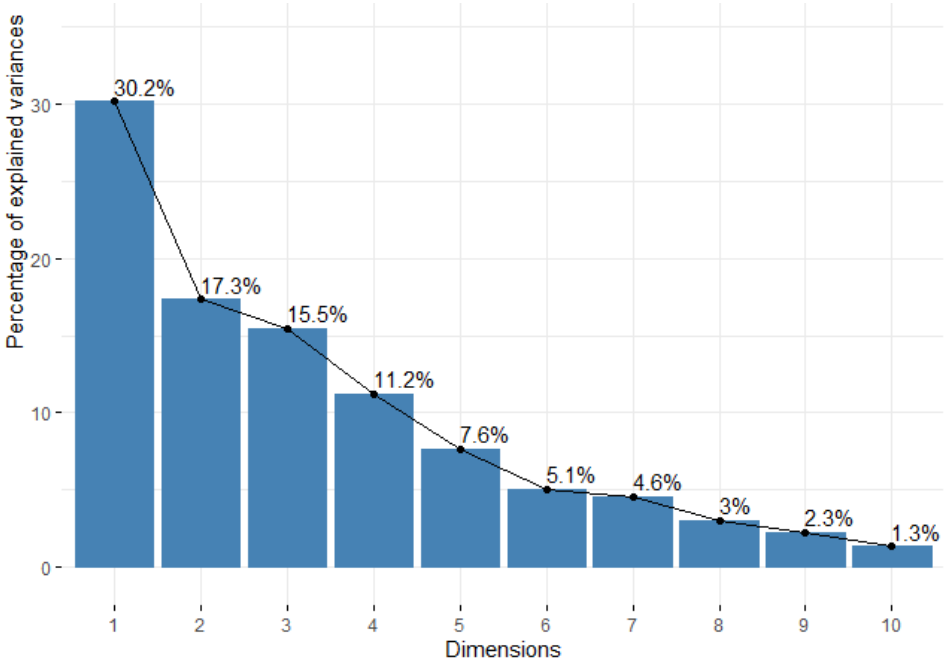
Dans un premier temps, nous allons analyser un modèle d'analyse en composantes principales. Contrairement à l'analyse univariée, l'analyse multidimensionnelle permet d'étudier les liens entre plusieurs variables et/ou plusieurs individus. Commençons par le choix de nombre de composantes principales à utiliser et pour cela on va se baser sur les méthodes existantes dans la littérature qui sont globalement trois solutions. les trois tois règles fréquemment utilisés sont les suivantes :

- 1ere regle : on regarde la valeur propre si la valeur est sup à 1 on le garde (Kaiser 1961).
- 2ème règle : « **méthode du coude** » qui consiste à repérer l'endroit à partir duquel le pourcentage d'inertie diminue beaucoup plus lentement lorsque l'on parcourt le diagramme des éboulis de gauche à droite.
- 3eme règle : on regarde la var cumulé en gardant les dimensions qui englobent une variance cumulés jusqu'à 80%.

Malheureusement, il n'existe pas de méthode objective bien acceptée pour décider du nombre d'axes principaux qui suffisent. Cela dépendra du domaine d'application spécifique et du jeu de données spécifiques.

Dans le graphique ci-dessus, nous pourrions vouloir nous arrêter à la cinquième composante principale. 81% des informations (variances) contenues dans les données sont conservées par les cinq premières composantes principales. On remarque que la première dimension représente la variance expliquée la plus élevée alors qu'elle résume plus que 30% de l'information et donc on peut proposer l'existence d'une forte corrélation entre les variables explicatives dans notre modèle.

Graphique 6 : Pourcentage de variances expliqués par dimension



Le tableau ci-dessous présente la même information mais cette fois-ci, nous allons nous concentrer sur les valeurs propres de chaque dimension et nous allons choisir les cinq premières dimensions puisqu'elles

respectent des valeurs propres supérieures à 1. Il faut bien noter que moins les variables sont corrélées plus la variance va être dispersée sur les différents axes. Nous pouvons déjà prédire l'existence d'une potentielle multicolinéarité entre les variables.

Tableau 1 : statistiques des variables numériques

	eigenvalue	**variance.percent**	**cumulative.variance.percent**
Dim.1	4.223776e+00	3.016983e+01	30.16983
Dim.2	2.427730e+00	1.734093e+01	47.51076
Dim.3	2.168340e+00	1.548815e+01	62.99890
Dim.4	1.565405e+00	1.118146e+01	74.18037
Dim.5	1.066863e+00	7.620447e+00	81.80081
Dim.6	7.101254e-01	5.072324e+00	86.87314
Dim.7	6.383454e-01	4.559610e+00	91.43275
Dim.8	4.245461e-01	3.032472e+00	94.46522
Dim.9	3.181069e-01	2.272192e+00	96.73741
Dim.10	1.858792e-01	1.327709e+00	98.06512
Dim.11	1.379387e-01	9.852763e-01	99.05040
Dim.12	8.230420e-02	5.878871e-01	99.63828
Dim.13	5.064019e-02	3.617156e-01	100.00000
Dim.14	1.494818e-31	1.067727e-30	100.00000

Passons à présent à l'analyse simultané du cercle de corrélations et le tableau représentant les contributions de chaque variable dans la construction des axes 1 et 2. L'examen du cercle de corrélation permet de détecter les éventuels groupes de variables qui se ressemblent ou au contraire qui s'opposent donnant ainsi un sens aux axes principaux. Dans un premier lieu Les variables sont représentées selon leurs corrélations avec les axes. La valeur de la corrélation dépend de la longueur de flesh. On remarque que les variables qui corrèlent le plus à la dimension 1 sont Abstentions_1 (une contribution de **0.77259540**), ROUSSEL(**0.57551672**), HIDALGO(**0.51770936**), POUTOU (**0.58475451**), Blancs_Nuls_1(**0.51281675**). Même les variables suivantes contribuent au concept représenté par ce axe 1 mais avec une contribution plus faibles : LASSALLE(**0.28587023**), ARTHAUD(**0.19870194**), MACRON(**0.21781387**), LE_PEN(**0.21529856**).
La dim2 est principalement corrélée avec les variables JADOT(**5.597950e-01**), MELENCHON**4.992486e-0**), LE_PEN(**3.663662e-01**) et ,ARTHAUD(0.57464**3.302213e-01**8806). Même les variables suivantes représentent partiellement le concept représenté par l'axe 2: DUPONT_AIGNAN(**2.699516e-01**),MACRON(**1.772729e-01**).
Pour les autres variables leurs coordonnées sur la cercle sont proches du centre donc elles sont moins importantes pour les premières composantes.En d'autres termes, plus la coordonnée d'une variable sera importante, plus la variable contribuera au concept représenté par ces axes.

Selon la représentation des variables sur la cercle de corrélations, on observe aussi :

- une corrélation négative très élevée entre les variables Abstentions_1 et HIDALGO => une colinéarité négative élevée.
- une colinéarité positive presque parfaite entre les variables LE_PEN et ARTHAUD. même constatation mais avec une colinéarité plus faible entre ces derniers variables et DUPONT. Elles peuvent presenter un groupe de variables qui varient ensemble.
- Meme constat avec les variables Blancs_Nuls_1, POUTOU,LASSALLE, PECRESSE et ROUSSEL.
- L'angle formé par les vecteurs colonnes renseignent la corrélation sur les variables
- Toutes les variables sont corrélées positivement sauf pour la variable Abstentions_1.
- Dans notre cas toutes les variables (sauf Abstentions_1) sont corrèlent positivement avec l'axe 1 alors on peut parler de l'Effet "Taille" c'est à dire que la 1er dimension va définir "un facteur de taille" ainsi les individus sont rangés sur l'axe 1 par valeurs croissantes de l'ensemble des variables (en moyenne).

Graphique 7 : Cercle de corrélations

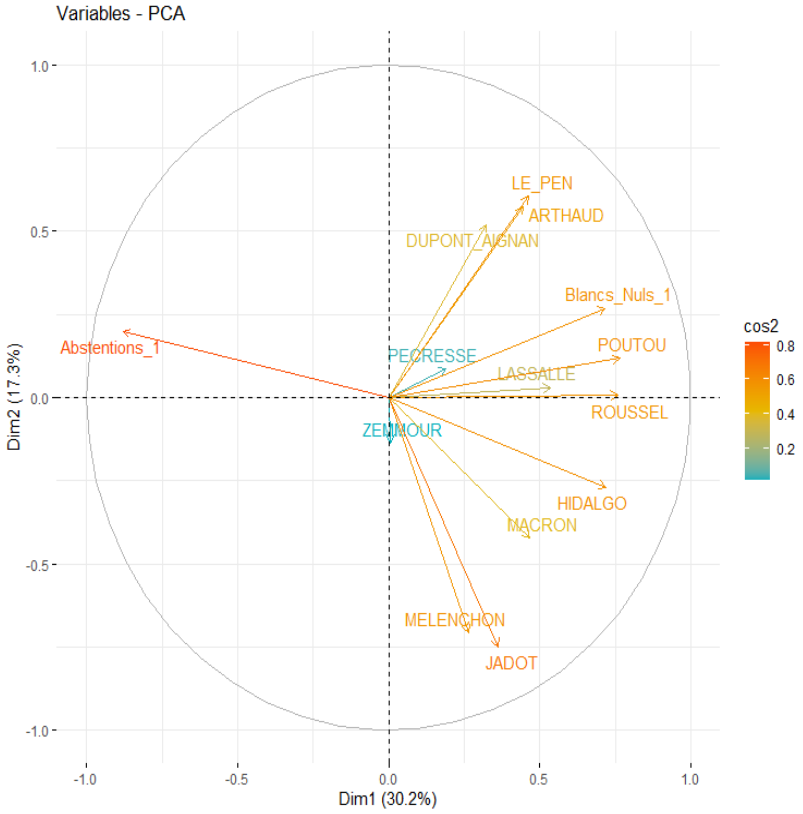
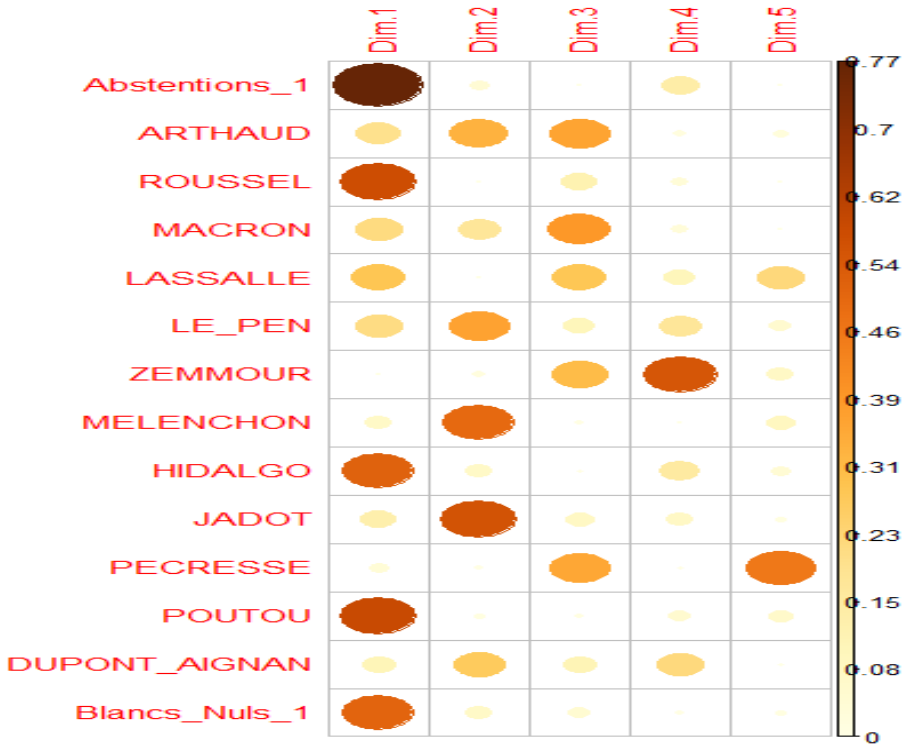


Tableau 2 :Contributions des variables dans la construction des axes

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
Abstentions_1	0.77259540	3.944654e-02	0.001335367	0.144198200	0.0001163684
ARTHAUD	0.19870194	3.302213e-01	0.359869401	0.015772340	0.0237829356
ROUSSEL	0.57551672	4.718568e-05	0.121467048	0.032528693	0.0006008429
MACRON	0.21781387	1.772729e-01	0.386963601	0.027924645	0.0020868454
LASSALLE	0.28587023	9.318359e-04	0.279227087	0.098906848	0.2246208662
LE_PEN	0.21529856	3.663662e-01	0.098500199	0.174146277	0.0535071510
ZEMMOUR	0.00003007	1.861669e-02	0.305357486	0.552348416	0.0762449560
MELENCHON	0.07023187	4.992486e-01	0.006936700	0.001269233	0.0885983781
HIDALGO	0.51770936	7.266865e-02	0.005143982	0.158390711	0.0402160337
JADOT	0.13202589	5.597950e-01	0.080885574	0.074822367	0.0132025345
PECRESSE	0.03605164	7.364491e-03	0.354774904	0.004924128	0.4610477112

POUTOU	0.58475451	1.392212e-02	0.009642336	0.052200282	0.0650577262
DUPONT_AIGNAN	0.10435938	2.699516e-01	0.111066416	0.220634436	0.0034390707
Blancs_Nuls_1	0.51281675	7.187684e-02	0.047170230	0.007338287	0.0143410915

Graphique 8 : Contributions des variables dans la construction des axes 1 et 2

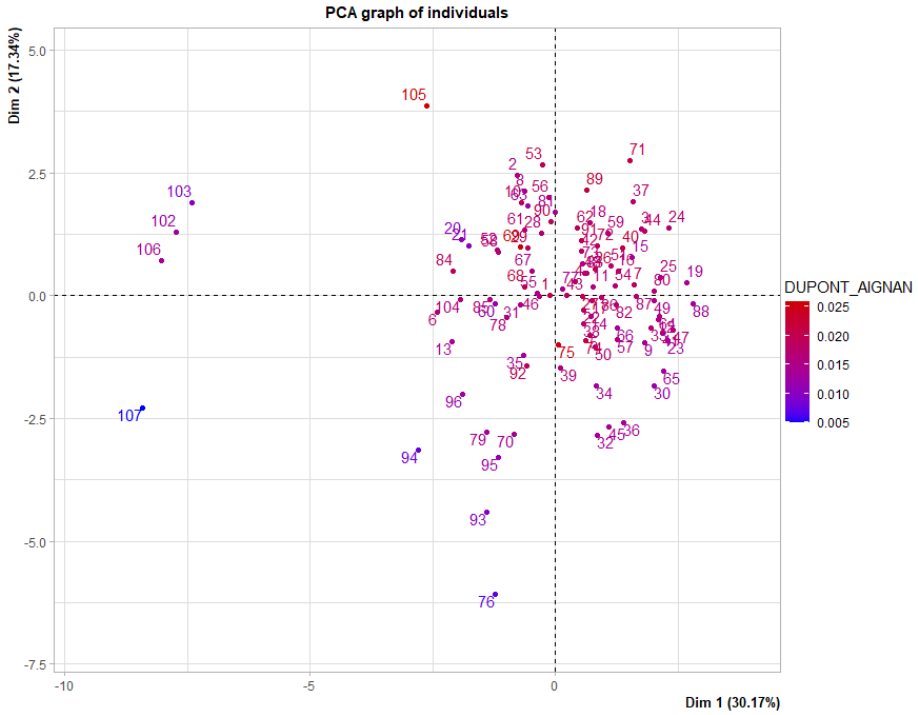


On s'intéresse maintenant aux individus et leurs représentations sur les axes. Le graphique ci-dessous affiche les dimensions et les coordonnées des individus : on regarde s' il y a des individus qui se rapprochent et on peut voir aussi la contribution de chaque individu sur les axes 1 et 2. Il faut noter que les individus qui sont proches de 0 n'ont pas beaucoup contribué à l'axe(comme les départements 43,77,46…). Aussi, on va regarder la qualité de représentation de chaque individu dans la construction des axes 1 et 2. Les individus qui ont une coordonnée élevée sur l'axe 1 sont caractérisés par un valeur élevée des variables qui sont corrélés significativement à cet axe.

Principalement, on cherche à détecter des individus opposés le but de déterminer de cluster. En d'autre terme, les individus proches le long d'une composante principale sont des individus qui partagent les mêmes caractéristiques vis-à-vis des variables quantitatives étudiées. Par exemple, les départements qui se présentent en haut à droite sont des départements ou on va trouver une proportion importante de vote répartie entre ROUSSEL, POUTOU,HIDALGO, LE PEN,LASSALLE, ARTHAUD et MACRON. Aussi ces départements vont avoir des proportions des Abstentions faibles. Dans notre cas on constate que nos individus présentent des coordonnées condensés autour de la moyenne au même temps on constate aussi des contributions excessive qui constitue un facteur d'instabilité comme dans le cas des individus 103, 102, 106(ces trois individus comportent presque de la même manière) de même les individus 76, 107,etc. Le retirer de ces derniers peut modifier

profondément le résultat de l'analyse. On a alors intérêt à effectuer l'analyse en éliminant puis à le rajouter et de comparer ensuite les résultats.

Graphique 9 :Les coordonnées des individus sur les axes 1 et 2



2.2 **Analyse multidimensionnelle de tour 2**

En se basant sur les mêmes règles présentées dans l'ACP de tour 1 pour le choix de nombre des axes, nous allons prendre en compte uniquement les deux premiers axes qui présentent plus que 80% (graphique 10) de la variance expliqués cumulés et au même temps des valeurs propres supérieurs à 1(Tableau 3)

Graphique 10 : Pourcentage de variances expliquées de chaque axes

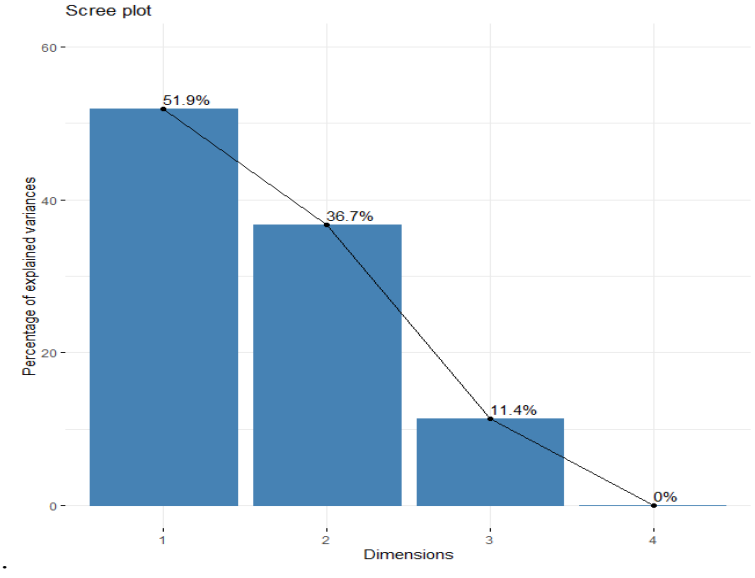


Tableau 3 : Le résultat de l'ACP sur les variables de tour 2

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	2.077728e+00	5.194319e+01	51.94319
Dim.2	1.467715e+00	3.669287e+01	88.63606
Dim.3	4.545575e-01	1.136394e+01	100.00000
Dim.4	3.277474e-30	8.193685e-29	100.00000

Ensuite, nous analysons simultanément la cercle de corrélations (graphique 11) et le tableau représentant les contributions (tableau 4) de chaque variable dans la construction des axes 1 et 2. D'après la cercle de corrélation, on illustre les conclusion suivantes :

- Une corrélation presque égale à -1 entre la variables Blancs_Nuls2 et Abstentions2.
- Même dans le cadre de tour 2, toutes les variables sont corrélées positivement avec l'axe 1 sauf pour la variable Abstentions_2.
- L'angle entre la variables LE_PEN_2 et MACRON_2 obtus et donc la correlations entre ces deux dernières va être négatif.
- La variable Blancs_Nuls_2 est corrélé positivement et quasi parfaitement avec la 1ere dimension et donc elle est la source principale de valeur de variances explique de cet axe. En second lieu on trouve la variables Abstentions_2 mais cette fois ci se présente dans la partie négative de l'axe avec une corrélation négative et très élevé avec l'axe. Cette illustration est justifiée par le tableau de contributions des variables sur les axes avec des valeurs égale à 0.9216041 pour Abstentions_2 et de 0.7088001 Blancs_Nuls_2.
- Même interprétation pour le deuxième axe qui est fortement corrélé positivement avec la variables LE_PEN_2 et négativement MACRON_2. Cette illustration est justifiée par le tableau de coordonnées des variables sur les axes avec des coordonnées très élevées.
- Les variables les plus corrélés à l'axe 2 sont LE_PEN_2 avec une contribution de 0.6247325243 et MACRON_2 avec une contribution de 0.8272628498. La Pen va avoir une corrélation positive avec l'axe et MACRON2 va avoir une corrélation positive avec ce dernier.
- On ne constate pas une moyenne corrélations entre les variables grâce aux angles qui séparent ces derniers. Une corrélation qui varie entre -0.6 et 0.3.

Graphique 11 : Cercle de corrélations

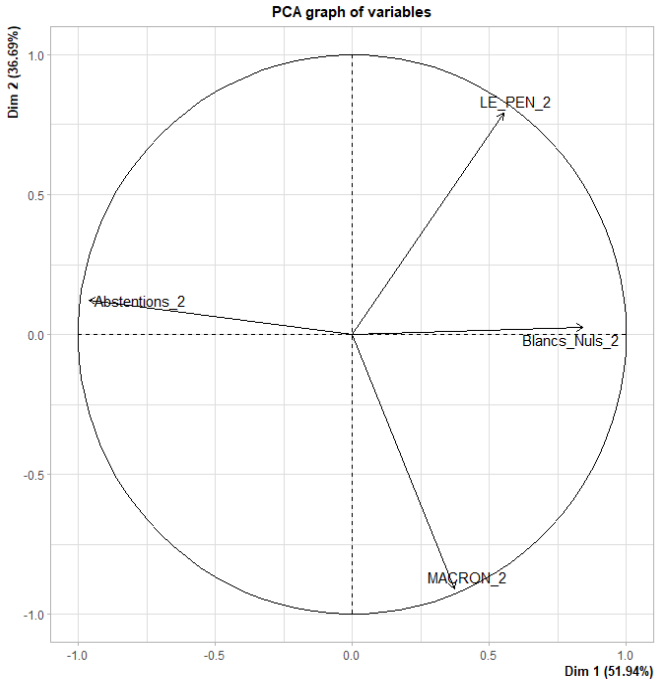
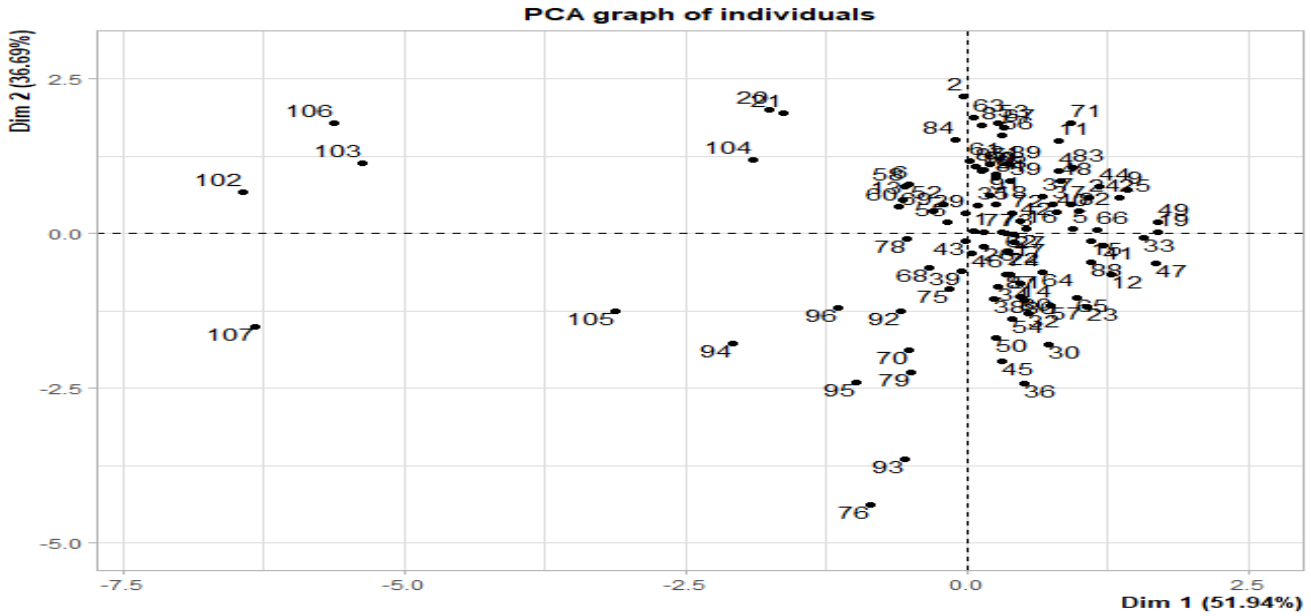


Tableau 4: Les contributions des variables dans les axes

	Dim.1	Dim.2	Dim.3	Dim.4
Abstentions_2	0.9216041	0.0150667614	0.06332918	1.333442e-30
MACRON_2	0.1407925	0.8272628498	0.03194467	9.777389e-31
LE_PEN_2	0.3065309	0.6247325243	0.06873655	9.032785e-31
Blancs_Nuls_2	0.7088001	0.0006528073	0.29054708	6.301460e-32

Le graphique ci-dessous affiche les coordonnées des individus sur les deux premiers axes.

Graphique 12 : Coordonnées des individus

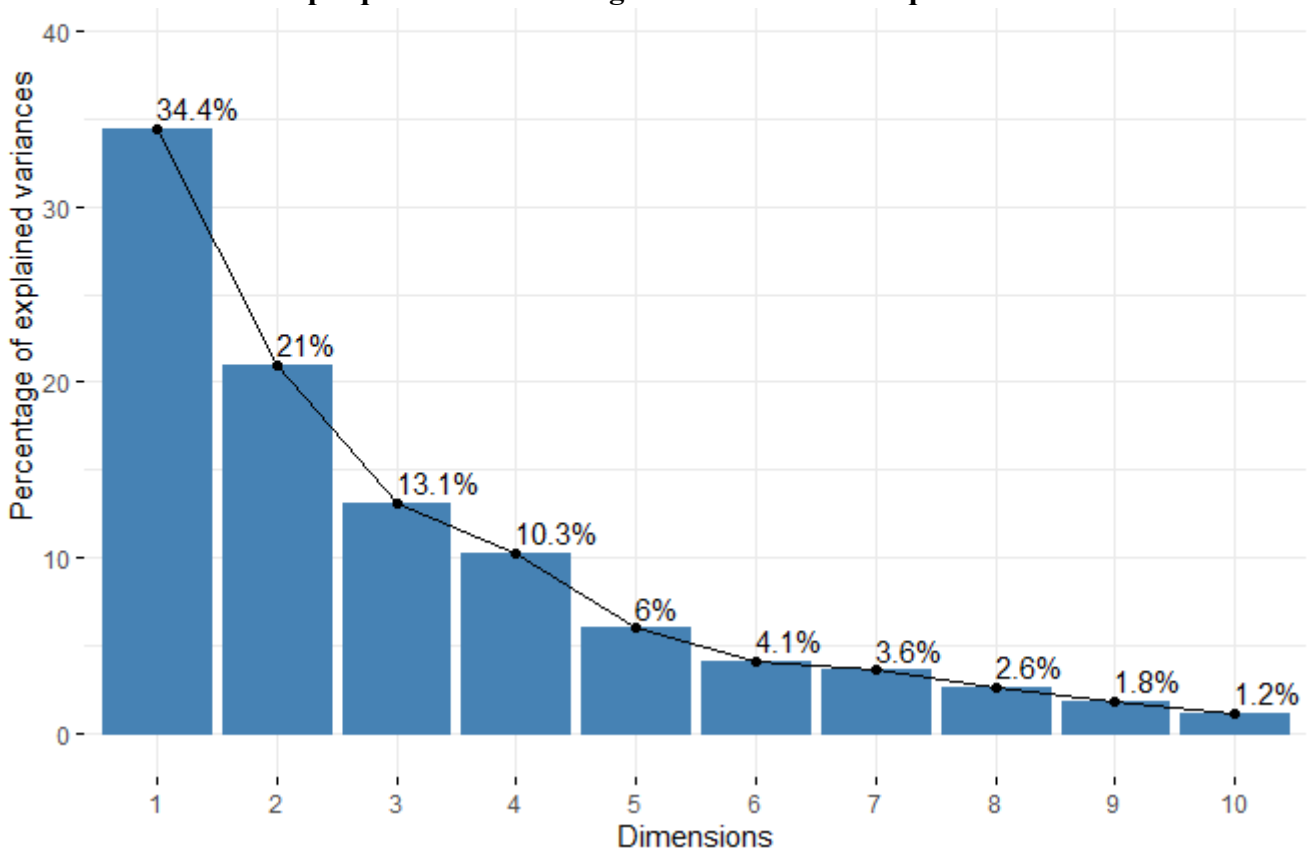


Les départements sont représentés presque sous la même forme que dans l'ACP sur le tour 1. La position des individus peut aussi être interprétée en relation avec la position des variables. Par exemple, les départements qui sont en haut à droite vont avoir tendance à voter pour LE_PEN et la partie en bas a tendance à voter pour MACRON. Ils sont aussi caractérisés par un taux d'abstention faible. Les départements qui ont les coordonnées les plus élevées sur l'axe 1 vont avoir un pourcentage considérable de Blancs et de Nuls.
Les départements qui se positionnent à gauche de l'axe 2 semblent avoir plus de pourcentage de abstentions.

2.3 Analyse multidimensionnelle de tour 1 et 2 simultanément

Les 5 premiers axes de l' ACP expriment presque 85% (graphique 13) de l'inertie totale du jeu de données ; cela signifie que 85% de la variabilité totale du nuage des individus (ou des variables) est représentée dans cette surface a 5 dimensions présentant des valeurs propres supérieurs à 1. On va admettre les 5 premiers axes dans le cadre de notre analyse.

Graphique 13 : Pourcentage de variance de chaque dimension



Le cercle de corrélation ci-dessous présente les corrélations entre les variables ainsi que la corrélation de ces derniers avec les deux axes 1 et 2. ce cercle résume les deux analyses précédentes et donc nous allons constater les mêmes interprétations en termes de corrélations et les sens de variations de celle ci.Par construction, le cosinus de l'angle de deux vecteurs variables permet de déterminer le coefficient de corrélation entre ces variables.

On constate ainsi (graphique 14) que le premier axe est corrélé de manière positive avec 14 variables initiales sauf les Absences1 et les Abstentions2 : plus un candidat aura une votation, plus les variables liées aux absentions auront des valeurs qui varient dans le sens inverse, c'est-à-dire qui sont décroissantes en fonction de pourcentage de vote. Tout cela est confirmé par le tableau de contributions qui montre les valeurs des contributions des variables dans la contributions de ces deux axes.

Les variables relatives à l'absentéisme et à l'abstention varient dans le sens inverse, c'est-à-dire qu'elles diminuent. Concernant l'axe 2, il oppose quant à lui, d'une part, les forts votes pour LE PEN pour les deux tours et MACRON (pour les deux tours) avec une corrélation négative entre eux. Il s'agit donc d'un axe d'opposition entre différentes visions politiques. On peut construire 4 groupes de variables en se basant sur les angles faibles qui existent entre elles.

groupe 1 : MACRON2, JADOT, MACRON, MELECHON et PECRESSE
groupe 2 : POUTOU, BLANC NULS2 et 1, ARTHAUD, ROUSSEL
groupe 3 : ZEMMOUR, DUPONT, LASSALLE, PEN 1 et 2
groupe 4 : Abstentions 1 et Abstentions 2

Cette interprétation peut être précisée avec les graphiques et tableaux relatifs aux individus que nous présentons maintenant.

Graphique 14 :Cercle de corrélations

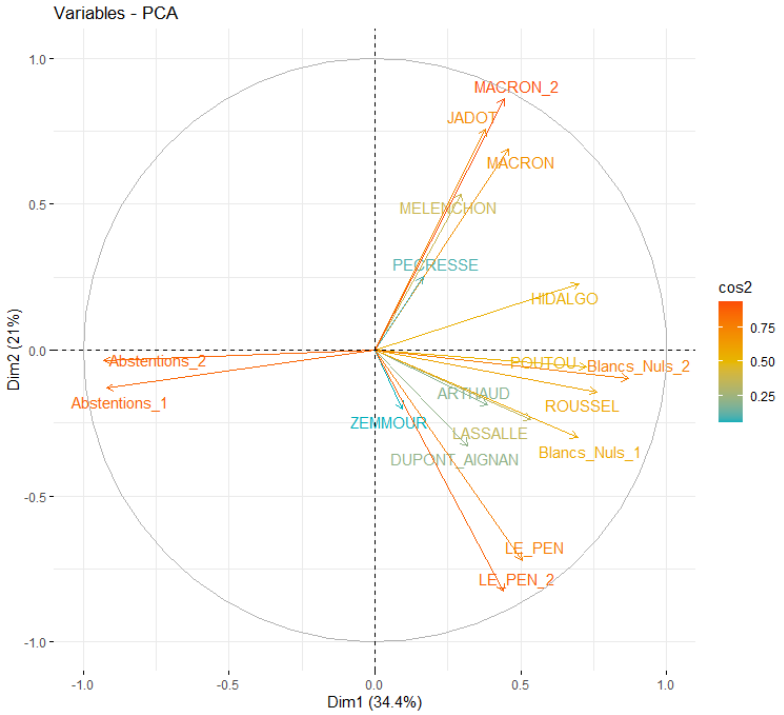
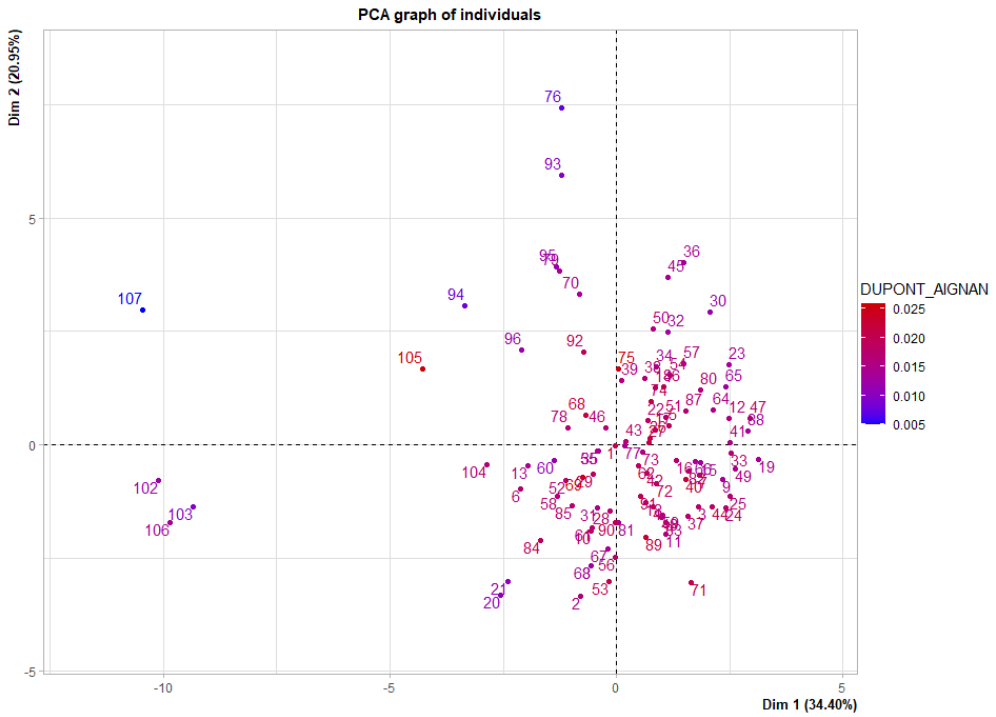


Tableau 5: Contributions des variables dans la constructions des axes

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
Abstentions_1	13.6377099	0.45329755	1.234675e-05	6.2355111	9.457194e-04
ARTHAUD	2.4201977	0.93762088	2.600750e+01	5.9072252	2.146900e+00
ROUSSEL	9.3080620	0.54130149	3.558888e+00	1.7976422	1.108042e-02
MACRON	3.3863210	12.59350126	6.892422e+00	0.8345958	4.192347e-02
LASSALLE	4.6190325	1.50793674	1.027914e+01	5.3961602	1.948589e+01
LE_PEN	4.0967563	13.73266376	9.499052e-01	6.3608047	3.056621e+00
ZEMMOUR	0.1459106	1.07224737	6.240304e+00	34.3293040	1.001216e+01
MELENCHON	1.4222302	7.59541691	6.719842e+00	1.1656723	9.665645e+00
HIDALGO	7.8630814	1.35025312	2.602191e+00	8.5701445	2.156979e+00
JADOT	2.3174680	15.18971813	8.674742e-02	5.1657536	1.680399e+00
PECRESSE	0.4351423	1.66714942	1.304698e+01	0.4393008	4.208006e+01
POUTOU	8.4699960	0.08949744	5.735789e-01	5.9214136	6.806377e+00
DUPONT_AIGNAN	1.6386995	2.83810184	1.581055e+01	2.8388425	1.200269e+00
Abstentions_2	13.9665670	0.03385206	7.252929e-01	4.6721952	1.451339e-02

MACRON_2	3.1497880	19.64138230	1.014389e+00	0.9483688	6.730313e-02
LE_PEN_2	3.1409398	18.12133992	3.371369e-01	4.3495773	3.037881e-01
Blancs_Nuls_1	7.7980388	2.39080160	1.016713e-01	1.8672914	1.008334e+00
Blancs_Nuls_2	12.1840590	0.24391824	5.053443e+00	3.2001968	2.608090e-01

Graphique 14 : Les coordonnées des départements sur les axes 1 et 2



Ce graphique montre les “ressemblances” entre individus. Plus ils sont proches (comme les individus 50, 32,57,etc en haut à droite), et plus leurs “profils” sont vraisemblablement similaires.
Comme dans le cas des variables, les individus les mieux représentés par la surface continent les cinq axes sont ceux les plus éloignés du centre. La degreadation de couleur permet de presneter l’intensite de valeur de cosinus autrement dit la qualite de representation sur les axes.

La position des individus peut aussi être interprétée en relation avec la position des variables. Par exemple, les departements tout à droite(comme 47,12,41,64…) , ont vraisemblablement eu des proportions assez élevées des variables POUTOU, Blancs_Nuls_1, ROUSSEL et LASSALLE. Les individus qui se positionnent à gauche de l’axe 2 semble avoir plus de pourcentage d'Abstentions dans le premier et le deuxième tour.

Les individus libellés sont ceux ayant la plus grande contribution à la construction du surface à 5 axes.

Nous pouvons valider l'existence d’une forte multicolinéarité entre les variables de 1er tour, 2eme tour et les variables de tour 1 et 2 entre eux. Donc, il faut chercher la méthode qui va nous donner la prédiction de proportions de variables de deuxième tour par celles de 1er tour tout en luttant contre l'existence de la corrélation entre les différentes variables explicatives.

3. Modélisation unidimensionnelle

L'analyse de la méthode des moindres carrés ordinaires (MCO) est l'une des techniques les plus couramment utilisées pour modéliser les relations entre les variables. Elle va nous permettre de mesurer la relation entre une variable dépendante et une ou plusieurs variables indépendantes, en minimisant la somme des carrés des écarts entre les valeurs observées et les valeurs prédites du modèle.

3.1 MCO

L'analyse de régression linéaire a pour objectif de prédire la variable dépendante "MACRON_2" à partir d'un ensemble de variables indépendantes, y compris les scores des candidats à l'élection présidentielle et des facteurs tels que les abstentions, les votes blancs ou nuls.

- **Premier modèle :** Modelisation de MACRON2 l'aide de tous les prédicteurs MCO

$$\textit{MACRON\_2} = \beta_0 + \beta_1 \text{ ARTHAUD} + \beta_2 \text{ ROUSSEL} + \beta_3 \text{ MACRON} + \beta_4 \text{ LASSALLE} + \beta_5 \text{ LE\_PEN} + \beta_6 \text{ MELENCHON} + \beta_7 \text{ HIDALGO} + \beta_8 \text{ JADOT} + \beta_9 \text{ PECRESSE} + \beta_{10} \text{ POUTOU} + \beta_{11} \text{ ZEMMOUR} + \beta_{12} \text{ DUPONT\_AIGNAN} + \beta_{13} \text{ Blancs\_Nuls\_1} + \beta_{14} \text{Abstentions\_1} + \varepsilon_1$$

Tableau 6: Sortie de résultats d'une régression linéaire multiple

	Estimate	Std.Error t	value	Pr(>|t|)	corr(y,x) simple
Abstentions_1	0.031502	0.007782	4.048	0.000111***	-0.22***
Blancs_Nuls_1	-0.986343	0.319447	-3.088	0.002698**	-0.08
ARTHAUD	4.967289	1.746706	2.844	0.005542**	0.1
ROUSSEL	-0.102084	0.215553	-0.474	0.636965	0.38***
MACRON	1.295697	0.036648	35.355	< 2e-16***	0.94***
LASSALLE	0.419880	0.088317	4.754	7.72e-06***	0.1
LE_PEN	-0.041018	0.035002	-1.172	0.244408	-0.05
ZEMMOUR	0.164124	0.110126	1.490	0.139713	0.23*
MELENCHON	0.541895	0.022429	24.160	< 2e-16***	0.29**
HIDALGO	-0.014943	0.297508	-0.050	0.960054	0.47***
JADOT	1.380570	0.179322	7.699	1.90e-11***	0.86***
PECRESSE	0.552489	0.095046	5.813	9.67e-08***	0.46***
POUTOU	-3.442229	0.994552	-3.461	0.000832***	0.37***
DUPONT_AIGNAN	-1.841430	0.321222	-5.733	1.37e-07**	0.20*

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.007091 on 88 degrees of freedom
Multiple R-squared: 0.9997, Adjusted R-squared: 0.9996
F-statistic: 2.046e+04 on 14 and 88 DF, p-value: < 2.2e-16
RMSE: 0.0000693574.

Ce modèle de régression MCO qui estime la proportion de vote pour Macron lors de deuxième tour à l’aide de l’ensemble des variables de tour 1 obtient un R2 presque parfait de 0.9997. Toutefois, plusieurs variables ne sont pas significatives, alors qu’elles l’étaient avec les corrélations simples et inversements. Comparé à l’analyse bivarié, on constate une différence entre les corrélations simples et multiples en termes de valeurs

sens et significativités. Donnons l'exemple des variables ARTHAUD et LASSALLE qui ont une corrélation multiple significative et élevées avec des valeurs respectives de 4.967289 et 0.419880. Par contre les corrélations simples de ces variables sont faibles et non significatives. On constate aussi une différence de signe et de valeur pour les variables POUTOU et Abstentions1. En effet, la variables POUTOU avec une corrélation simple positive à une corrélation de régression multiple négative et inversement, pour la variable Abstentions 1. Pour la variables HIDALGO, on remarque une augmentation de valeur, de significativité et un changement de signe en passant de la régression multiple a la corrélation simple c'est à dire en passant de -0.014943 a 0.47***.

Tableau 7 : Coefficients des prédicteurs

	Beta MCO
Abstentions_1	0.03150187
Blancs_Nuls_1	-0.98634251
ARTHAUD	4.96728895
ROUSSEL	-0.10208363
MACRON	1.29569703
LASSALLE	0.41987973
LE_PEN	-0.04101779
ZEMMOUR	0.16412410
MELENCHON	0.54189470
HIDALGO	-0.01494338
JADOT	1.38056962
PECRESSE	0.55248945
POUTOU	-3.44222900
DUPONT_AIGNAN	-1.84142979

- **Deuxième modèle:** Modélisation de LE_PEN2 l'aide de tous les prédicteurs MCO

LE_PEN2= β_0 + β_1 ARTHAUD + β_2 ROUSSEL + β_3 MACRON + β_4 LASSALLE + β_5 LE_PEN + β_6 MELENCHON + β_7 HIDALGO +β_8 JADOT + β_9 PECRESSE + β_{10} POUTOU + β_{11} ZEMMOUR + β_{12} DUPONT_AIGNAN + β_{13} Blancs_Nuls_1 + β_{14}Abstentions_1 + ε_1
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Tableau 8: Sortie de résultats d'une régression linéaire multiple

	Estimate	Std.Error t	value	Pr(>|t|)	corr simple
Abstentions_1	0.09452	0.01122	8.424	6.28e-13***	-0.14***
Blancs_Nuls_1	0.20987	0.46060	0.456	0.649764	0.41***
ARTHAUD	0.52912	2.51851	0.210	0.834082	0.32***
ROUSSEL	-0.36310	0.31080	-1.168	0.245847	0.37***
MACRON	-0.22118	0.05284	-4.186	6.72e-05***	-0.17
LASSALLE	0.22330	0.12734	1.754	0.082988	0.29**
LE_PEN	1.16945	0.05047	23.172	< 2e-16***	0.92**
ZEMMOUR	0.62791	0.15879	3.954	0.000155***	0.36***
MELENCHON	0.04483	0.03234	1.386	0.169224	-0.27**

HIDALGO	0.84860	0.42897	1.978	0.051028	0.04
JADOT	-0.35508	0.25856	-1.373	0.173150	-0.28**
PECRESSE	0.37183	0.13704	2.713	0.008016**	-0.11
POUTOU	4.32216	1.43401	3.014	0.003367**	0.30**
DUPONT_AIGNAN	1.04593	0.46316	2.258	0.026402*	0.48***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01022 on 88 degrees of freedom
Multiple R-squared: 0.999, Adjusted R-squared: 0.9988
F-statistic: 6216 on 14 and 88 DF, p-value: < 2.2e-16
RMSE: 0.0001407813

Notre modèle de régression linéaire multiple vise à prédire le score de Marine Le Pen (LE_PEN2) lors d'une élection en utilisant plusieurs prédicteurs (variables indépendantes de 1er tour). Le modèle dans son ensemble explique bien la variance dans les scores de LE_PEN2, avec un R carré ajusté de 0,9755, ce qui signifie que les variables indépendantes expliquent environ 99,99% de la variance dans les scores de LE_PEN2. Toutefois seulement la moitié de ces variables sont significatives bien que Le F-statistic de 6216 présente une p-value inférieur à 0 (< 2.2e-16) alors elle suggère que le modèle est statistiquement significatif. Comparons encore une fois l'analyse bivariée et la multivariée, on constate une différence entre les corrélations simples et multiples en termes de valeurs sens et significativités.

Les résultats indiquent que les variables ZEMMOUR, Abstentions_1, PECRESSE,POUTOU et MACRON sont significativement associées avec le score de LE_PEN2, car elles ont des valeurs de p inférieures à 0,05. En particulier, une unité de changement dans le score de MACRON est associée à une diminution de 0.22118 dans le score de LE_PEN2 . En revanche, les variables ARTHAUD, HIDALGO ne sont pas significativement associées avec le score de LE_PEN2. Alors que dans le cadre de la simple corrélation on constate que même les variables Blancs_Nuls et ARTHUAUD, LASSALLE sont très significatives. On constate aussi un changement de signe de corrélations et significations pour la variable MELENCHON en passant d'une corrélation simple négative et significative(une corrélation égale à -0.27**) avec une correction positive, faible et non significative. On constate la meme chose pour la variable Abstentions1.

Nous pouvons conclure que plusieurs variables ont perdu leurs significativités en passant d'une corrélation simple vers une corrélation multiple dans le cadre d'une régression multivariée.

Tableau 9 : Coefficients des prédicteurs

	Beta_MCO
Abstentions_1	0.09452174
Blancs_Nuls_1	0.20987149
ARTHAUD	0.52911827
ROUSSEL	-0.36309998
MACRON	-0.22117603
LASSALLE	0.22329947
LE_PEN	1.16944757
ZEMMOUR	0.62790860
MELENCHON	0.04482591
HIDALGO	0.84860309
JADOT	-0.35507648
PECRESSE	0.37182509
POUTOU	4.32216340

DUPONT_AIGNAN	1.04593368

- **Troisième modèle:** Modélisation de Abstentions_2 l'aide de tous les prédicteurs MCO

Abstentions_2 = β_0 + β_1 ARTHAUD + β_2 ROUSSEL + β_3 MACRON + β_4 LASSALLE + β_5 LE_PEN + β_6 MELENCHON + β_7 HIDALGO +β_8 JADOT + β_9 PECRESSE + β_{10} POUTOU + β_{11} ZEMMOUR + β_{12} DUPONT_AIGNAN + β_{13} Blancs_Nuls_1 + β_{14}Abstentions_1 + ε_1

Tableau 10: Sortie de résultats d'une régression linéaire multiple

	Estimate	Std.Error	t value	Pr(>|t|)	corr simple
Abstentions_1	0.88567	0.01192	74.311	< 2e-16***	-0.19***
Blancs_Nuls_1	0.59609	0.48925	1.218	0.2263	0.32**
ARTHAUD	-2.78580	2.67517	-1.041	0.3006	-0.44**
ROUSSEL	0.62656	0.33013	1.898	0.0610	-0.23**
MACRON	-0.02357	0.05613	-0.420	0.6756	0.082**
LASSALLE	0.12785	0.13526	0.945	0.3471	-0.42**
LE_PEN	-0.12994	0.05361	-2.424	0.0174*	0.64**
ZEMMOUR	0.28588	0.16866	1.695	0.0936	0.02**
MELENCHON	0.33930	0.03435	9.877	6.42e-16***	0.24
HIDALGO	-0.70299	0.45565	-1.543	0.1265	-0.22**
JADOT	-0.07737	0.27464	-0.282	0.7788	0.23**
PECRESSE	0.04913	0.14557	0.337	0.7366	-0.21**
POUTOU	-1.92662	1.52321	-1.265	0.2093	-0.44**
DUPONT_AIGNAN	1.02094	0.49197	2.075	0.0409*	-0.45**

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01086 on 88 degrees of freedom
Multiple R-squared: 0.9987, Adjusted R-squared: 0.9985
F-statistic: 4937 on 14 and 88 DF, p-value: < 2.2e-16
RMSE: 0.0001443698

Le modèle a un R carré ajusté de 0,9812, ce qui indique que le modèle explique bien la variation du taux d'abstention. Le modèle est également significatif avec une p-value très faible (< 2,2e-16), ce qui suggère que le modèle dans son ensemble est utile pour prédire le taux d'abstention.

Dans ce modèle bien que R2 est très grand, il n’existe que 4 variables qui présentent des coefficients significatifs. Nous pouvons conclure que la quasi-totalité des variables ont perdu leurs significations en passant d’une corrélation simple vers une corrélation multiple dans le cadre d’une régression multivariée et inversement. Par exemple les variables HIDALGO, JADOT, LASSALLE etc. On constate aussi plusieurs changements de sens de variations en régression toutes les variables ensemble dans le même modèle. Donnant l’exemple des variables LE_PEN, Abstentions 2 et PECRESSE. Par exemple, les votes blancs et nuls n'ont pas d'effet significatif sur le taux d'abstention au second tour. ce qui est inversement constaté dans le cadre de la corrélation simple en termes de significativité.

Tableau 11 : Coefficients des prédicteurs

	Beta_MCO
Abstentions_1	0.88566557
Blancs_Nuls_1	0.59608742
ARTHAUD	-2.78579582
ROUSSEL	0.62655860
MACRON	-0.02356905
LASSALLE	0.12785250
LE_PEN	-0.12994493
ZEMMOUR	0.28587720
MELENCHON	0.33929836
HIDALGO	-0.70298927
JADOT	-0.07737149
PECRESSE	0.04912708
POUTOU	-1.92662176
DUPONT_AIGNAN	1.02094227

- **Quatrième modèle:** Modélisation de Blancs_Nuls_2 l'aide de tous les prédicteurs MCO

Blancs_Nuls_2 = β_0 + β_1 ARTHAUD + β_2 ROUSSEL + β_3 MACRON + β_4 LASSALLE + β_5 LE_PEN + β_6 MELENCHON + β_7 HIDALGO +β_8 JADOT + β_9 PECRESSE + β_{10} POUTOU + β_{11} ZEMMOUR + β_{12} DUPONT_AIGNAN + β_{13} Blancs_Nuls_1 + β_{14} Abstentions_1 + ε_1

Tableau 12: Sortie de résultats d'une régression linéaire multiple

	Estimate	Std.Error	t value	Pr(>|t|)	corr simple
Abstentions_1	-0.0135390	0.0044525	-3.041	0.00311**	-0.16**
Blancs_Nuls_1	1.2211117	0.1827728	6.681	2.07e-09***	0,61**
ARTHAUD	-1.6959425	0.9993857	-1.697	0.09323	0.16
ROUSSEL	0.8332651	0.1233294	6.756	1.47e-09***	0.78***
MACRON	-0.0503220	0.0209683	-2.400	0.01851*	0.23*
LASSALLE	0.2272968	0.0505309	4.498	2.08e-05***	0.76***
LE_PEN	-0.0005998	0.0200264	-0.030	0.97617	0.36***
ZEMMOUR	-0.0731548	0.0630091	-1.161	0.24877	0.16
MELENCHON	0.0760231	0.0128330	5.924	5.98e-08***	0.29**
HIDALGO	0.8537414	0.1702206	5.016	2.72e-06***	0.78***
JADOT	0.0396379	0.1026001	0.386	0.70018	0.27**
PECRESSE	0.0233142	0.0543810	0.429	0.66917	0.05

POUTOU	2.0472883	0.5690375	3.598	0.00053***	0.67***
DUPONT_AIGNAN	0.7915673	0.1837885	4.307	4.29e-05***	0.21*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.004057 on 88 degrees of freedom
Multiple R-squared: 0.997, Adjusted R-squared: 0.9966
F-statistic: 2112 on 14 and 88 DF, p-value: < 2.2e-16
RMSE: 0.00002014837

La régression de la variables Blancs_Nuls_2 sur les 14 variables de tour 1 met en évidence des différences en termes de signe, de valeur et significativités entres plusieurs corrélations multiples et simples. On constate initialement que seulement 9 variables explicatives sont significatives bien que Le F-statistic de 2112 présente une p-value inférieur à 0 (< 2.2e-16) alors elle suggère que le modèle est statistiquement significatif dans toutes ses variables. En effet, il existe des corrélations simples positives qui sont devenues négatives dans le cadre de notre modèle en gardant la même significativité comme pour la variables MACRON. Des autres variables ont perdu en termes de significativité en passant de la cas simple vers un modèle multiple comme les variables LE_PEN (avec un inversement de signe) et JADOT. Toutefois, La variable DUPONT_AIGNAN a gagné en termes de significativité et elle a maintenant un estimateur très significatif.

Nous pouvons conclure que la cause principale de cette differences entre les corrélations multiples et la corrélations simples entre les variables à expliquer et les variables explicatives est dû à la multicolinéarité qui peut impacter la statistique de test de Student et par conséquences les p value ce qui peut expliquer les changements de significations d'une correlations d'une autre entre les mêmes variables.

Tableau 13 : Coefficients des prédicteurs

	Beta_MCO
Abstentions_1	-0.0135390230
Blancs_Nuls_1	1.2211117455
ARTHAUD	-1.6959425070
ROUSSEL	0.8332651109
MACRON	-0.0503219722
LASSALLE	0.2272967688
LE_PEN	-0.0005998119
ZEMMOUR	-0.0731547965
MELENCHON	0.0760231076
HIDALGO	0.8537413693
JADOT	0.0396379409
PECRESSE	0.0233142264

POUTOU	2.0472883358
DUPONT_AIGNAN	0.7915672526

- **Détection de la multicolinéarité (Variance Inflation Factor):**

Afin de détecter et de confirmer nos soupçons de multicolinéarité nous avons calculé la variance inflation factor entre nos 14 variables explicatives sur nos modèles de régression MCO. Le VIF mesure le degré auquel la variance d'un coefficient de régression estimé est augmentée en raison de la colinéarité dans le modèle. Un VIF de 1 indique l'absence de corrélation entre la variable indépendante et les autres variables, tandis qu'un VIF supérieur à 1 suggère une corrélation avec les autres variables. Généralement, une valeur VIF de 100 ou plus indique une forte multicolinéarité entre les variables.

En examinant nos résultats, nous pouvons constater que les valeurs VIF de deux variables est supérieures à 100, ce qui indique que la multicolinéarité est forte. Cependant, le reste des variables ont des valeurs VIF inférieur à 100, comme Abstentions_1, ROUSSEL ou PECRESSE, ce qui suggère qu'il ne peut pas y avoir de une très forte multicolinéarité entre ces variables et les autres dans le modèle.

Il est important de noter que la présence de multicolinéarité peut affecter la précision et l'interprétabilité de notre modèle de régression, il est donc important de traiter ce problème avant de poursuivre notre analyse. Nous pouvons envisager d'abandonner une ou plusieurs des variables fortement corrélées ou d'utiliser des techniques telles que la régression en composantes principales (PCR) ou de la régression PLS (Partial Least Squares) qui peuvent être des alternatives efficaces pour traiter le problème de la multicolinéarité.

Tableau 14: Résultats du VIF

Variable	VIF
Abstentions_1	9.163294
Blancs_Nuls_1	63.583412
ARTHAUD	131.252477
ROUSSEL	32.145223
MACRON	108.279424
LASSALLE	18.639462
LE_PEN	86.320211
ZEMMOUR	67.570362
MELENCHON	23.390143

HIDALGO	35.598276
JADOT	68.040305
PECRESSE	28.230229
POUTOU	73.957258

DUPONT_AIGNAN	57.335209

3.2 La régression sur composantes principales (PCR)

Nous allons maintenant appliquer une deuxième approche pour lutter contre la multicolorarite. La régression en composantes principales (RCP) permet tout d'abord de procéder à une analyse en composantes principales (ACP) sur les variables explicatives, qui sont généralement bruitées. On effectue ensuite une régression sur les composantes principales retenues. Dans le cadre de notre régression nous n'allons pas normaliser les données puisque nous n'avons pas de contrainte d'échelle et les valeurs de toutes les variables varient entre 0 et 1. Nous appliquerons le PCR sur les quatre modèles précédents.

- **Premier modèle : modèle *Modélisation de LE_PEN_2 l'aide de tous les prédicteurs de premier tour***

Dans le cadre du choix de composantes principales qui vont être les niveaux prédicteurs on a choisi les composantes significatives après la régression de la variables LE_PEN_2 sur toutes les composantes principales de L'ACP sur les variables de tour 1. on a fixé un seuil de 2.5% au delà de ce valeur les composantes principales ne sont pas significatives.

Tableau 15: RMSE er R2 de la régression sur composantes principales

	RMSE	R2
Régression sur toutes les variables	0.0001407813	0.999
Régression sur toutes les composantes	0.0001279564	0.999
Régression sur Z1, Z6 et Z14	0.001958971	0.5811

Le tableau ci-dessus, montre que on constate une grande différence en R2 en passant de la régression multiple sur toutes les variables vers la régression multiple sur les composantes les plus significatives ceci met en évidence la perte de l'information causé par l'élimination de beaucoup de composantes .
Le R2 résultant avec ces trois composantes sélectionnées perd évidemment un beaucoup. Il passe de 0.999 avec toutes les variables à 0.58116, avec seulement trois composantes principales. Aussi une faible perte en RMSE : 0.0001279564 vs 0.001958971.

Graphique 16 : Comparaison entre les coef de PCR et les corrélations simple

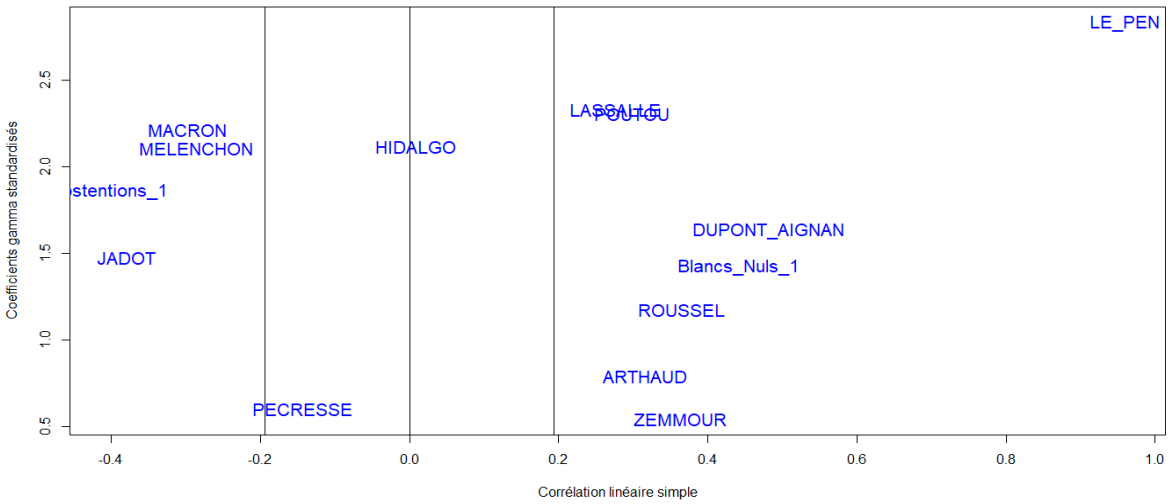


Tableau 16 : Coefficients des prédicteurs

Colonne1	**gamm_PCR**	corr simple
Abstentions_1	1.8592356	-0.14***
ARTHAUD	0.7902248	0.32***
ROUSSEL	1.1712853	0.37***
MACRON	2.2106592	-0.17
LASSALLE	2.3305518	0.29**
LE_PEN	2.8296652	0.92**
ZEMMOUR	0.5426811	0.36***
MELENCHON	2.1065524	-0.27**
HIDALGO	2.1135011	0.04
JADOT	1.4751500	-0.28**
PECRESSE	0.6008762	-0.11
POUTOU	2.3058307	0.30**
DUPONT_AIGNAN	1.6286894	0.41***
Blancs_Nuls_1	1.4231359	0.41***

Le tableau et le graphique 16 montrent que les variables qui agissent le plus et positivement sur PEN_2 sont MACRON, LASSALLEL E_PEN, MELENCHON, HIDALGO et POUTOU. Malheureusement, on ne peut pas réaliser des tests statistiques sur ces estimateurs car ils sont biaisés. Toutefois, on va comparer ces estimateurs aux signes des coefficients de corrélation simple entre chaque variable et LE_PEN2. Nous pouvons remarquer que les signes de Abstentions_1, MACRON, MELENCHON, JADOT ne sont pas en accord. En se basant aussi sur le graphique 16, on s'intéresse aux variables qui se trouvent dans la partie droite puisque tous les coef gamma sont positifs. Ce graphique vérifie notre conclusion par rapport aux coefficients. Ces changements de signes menacent la fiabilité et la puissance prédictive de notre modèle. Le fait de trouver que Abstentions1 augmente de 1.86 point de pourcentage avec l'augmentation d'une unité la proportion de vote pour PEN2 présente une alerte.

- **Deuxième modèle : modèle *Modélisation de MACRON_2 l'aide de tous les prédicteurs de premier tour***

Dans le cadre de choix de composantes principales qui vont être les niveaux prédicteurs on a choisi les composantes significatives après la régression de la variables *MACRON_2* sur toutes les composantes

principales de L'ACP sur les variables de tour 1. on a fixé un seuil de 2.5% au delà de ce valeur les composantes principales ne sont pas significatives.

Tableau 17: RMSE er R2 de la régression sur composantes principales

	RMSE	R2
Régression sur toutes les variables	0.0000693574	0.9997
Régression sur toutes les composantes	6.154793e-05	0.9997
Régression sur Z3,Z6 ,Z8 , Z10, Z12 , Z13 et Z14	8.954271e-05	0.984

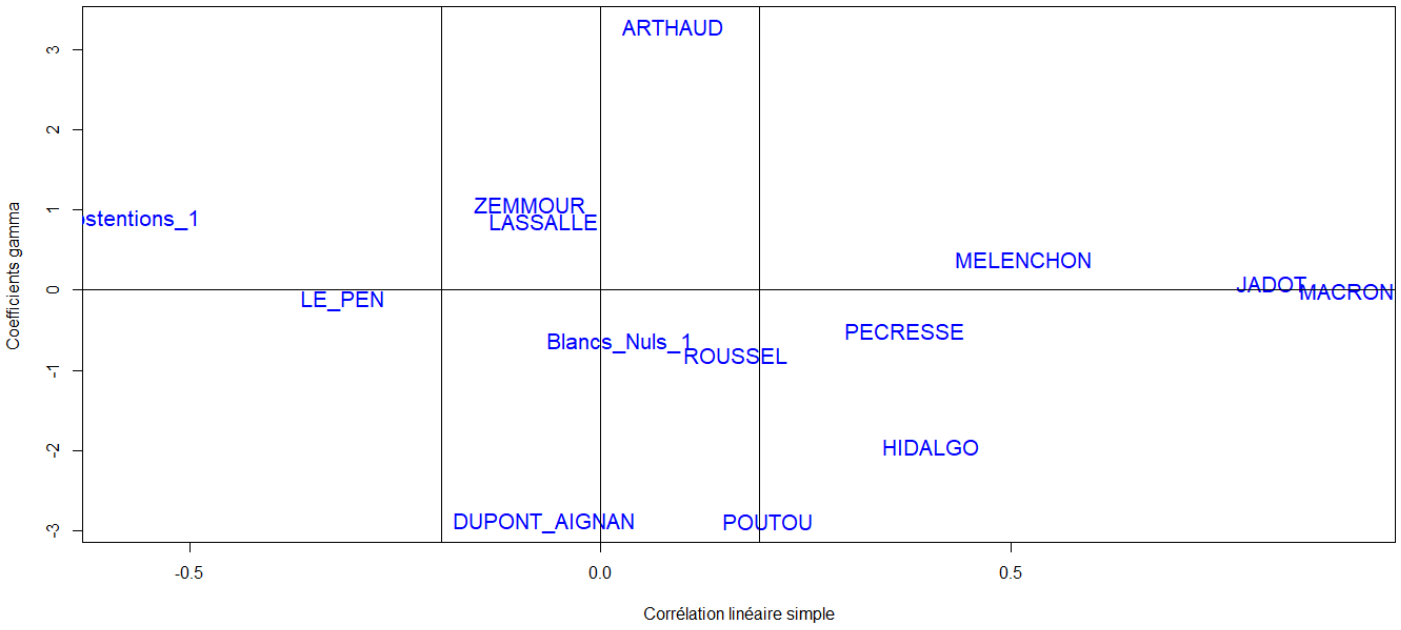
Le tableau ci-dessus, montre que si on utilise toutes les composantes principales, la méthode de régression sur composantes principales est équivalente à la méthode de régression multiple. Le R2 résultant avec ces septs composantes sélectionnées perd évidemment un peu. Il passe de 0.9997 avec toutes les variables à 0.984, avec seulement trois composantes principales. Un perd sur RMSE est plus élevé : 0.0000693574 vs 6.154793e-05.

Le tableau et le graphique 17 montrent que les variables qui agissent le plus en valeur absolue sur *MACRON_2* sont MACRON, HIDALGO ,DUPONT_AIGNAN et surtout ARTHAUD et POUTOU. Mais, ca sera impossible de réaliser des tests statistiques sur ces estimateurs car ils sont biaisés. Cependant, on va comparer ces estimateurs aux signes des coefficients de corrélation simple entre chaque variable et *MACRON_2*. Nous constatons que les signes de HIDALGO ,DUPONT_AIGNAN, POUTOU,PECRESSE, MACRON, ROUSSEL et Abstentions_1 ne sont pas en accord.Alors en passant par PCR on va se trouver un grand changement de signe de prédicteurs ce qui poussera des problèmes en terme d'interprétabilité des coefs en terme politique. En se basant aussi sur le graphique 17, on s'intéresse aux variables qui se trouvent dans la partie en haut à droite et la partie en bas à gauche puisque on cherche l'egalité de signes entre les gamma et les corrélations simples. Ce graphique appuie notre conclusion par rapport aux coefficients. Ces changements de signes menacent encore une fois la fiabilité et la puissance prédictive de notre modèle. Une autre alerte sous forme d'un coef gamma positive de Abstentions1.

Tableau 18 : Coefficients des prédicteurs

	gamm_PCR	corr(y,x) simple
Abstentions_1	0.88869841	-0.22***
ARTHAUD	3.28505396	0.1
ROUSSEL	-0.80261621	0.38***
MACRON	-0.01128135	0.94***
LASSALLE	0.85160509	0.1
LE_PEN	-0.12452710	-0.05
ZEMMOUR	1.07008830	0.23*
MELENCHON	0.38336681	0.29**
HIDALGO	-1.94520632	0.47***
JADOT	0.08423118	0.86***
PECRESSE	-0.50087328	0.46***
POUTOU	-2.88413294	0.37***
DUPONT_AIGNAN	-2.89162065	0.20*
Blancs_Nuls_1	-0.64491176	-0.08

Graphique 17 : Comparaison entre les coef de PCR et les corrélations simple



- **Troisième modèle** : modèle *Modélisation de* Blancs_Nuls_2 *l'aide de tous les prédicteurs de premier tour*

Dans le cadre de choix de composantes principales qui vont être les niveaux prédicteurs on a choisi les composantes significatives après la régression de la variables Blancs_Nuls_2 sur toutes les composantes principales de L'ACP sur les variables de tour 1. on a fixé un seuil de 2.5% au delà de ce valeur les composantes principales ne sont pas significatives.

Tableau 19: RMSE er R2 de la régression sur composantes principales

	RMSE	R2
Régression sur toutes les variables	0.00002014837	0.997
Régression sur toutes les composantes	2.014837e-05	0.997
Régression sur Z1, Z3, Z4, Z6 ,Z8 , Z10, Z12 et Z14	2.626683e-05	0.9289

Le tableau ci-dessus présente encore une fois la même conclusion. En effet, il montre que si on utilise toutes les composantes principales, la méthode de régression sur composantes principales est équivalente à la méthode de régression multiple.

En passant par le PCR, le R2 résultant après la régression de Blancs_Nuls_2 sur les 8 composantes sélectionnées perd en termes de pouvoir explicative a cause de la perte en termes de quantité d'information dans le modèle. Il passe de 0.9289 avec toutes les variables à 0.984, avec seulement 8 composantes principales. Un perd sur RMSE est plus élevé : 2.014837e-05 vs 2.626683e-05.

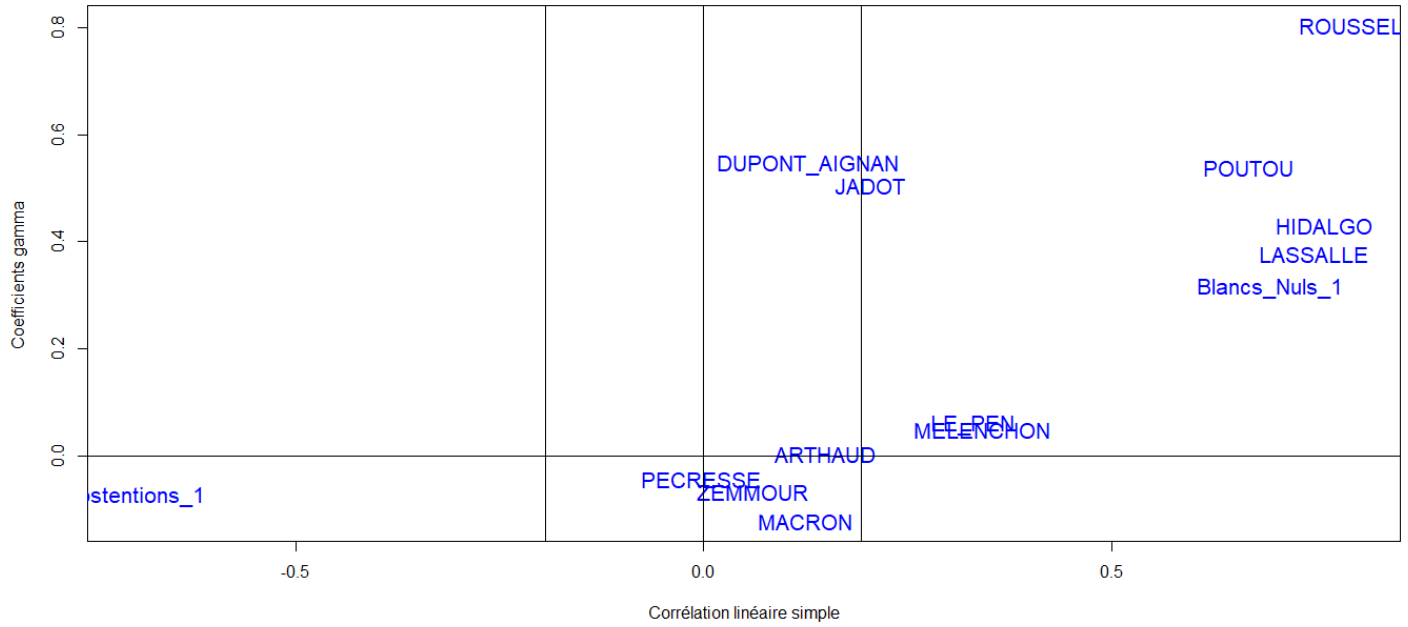
Tableau 20 : Coefficients des prédicteurs

	gamm_PCR	corr simple
Abstentions_1	-0.07554219	-0.16**
ARTHAUD	0.00362356	0.16

ROUSSEL	0.80443349	0.78***
MACRON	-0.12248514	0.23*
LASSALLE	0.37726146	0.76***
LE_PEN	0.05762821	0.36***
ZEMMOUR	-0.06833560	0.16
MELENCHON	0.04793250	0.29**
HIDALGO	0.42976884	0.78***
JADOT	0.50486436	0.27**
PECRESSE	-0.04413244	0.05
POUTOU	0.53938866	0.67***
DUPONT_AIGNAN	0.54530423	0.21*
Blancs_Nuls_1	0.31474159	0.61**

Le tableau et le graphique 18 illustrent les variables qui agissent le plus en valeur absolue sur Blancs_Nuls_2 sont ROUSSEL, DUPONT_AIGNAN ,MELENCHON ,HIDALGO, JADOT et POUTOU. On trouve le probleme pour realiser des tests statistiques sur les gamma. alors, on va comparer ces estimateurs aux signes des coefficients de corrélation simple entre chaque variable et Blancs_Nuls_2. Nous pouvons remarquer qu'il y a peu de coef gamma qui ont sebuir le changement de signe : MACRON, ZEMMOUR, PECRESSE. En s'appuyant sur le graphique 18, on s'intéresse aux variables qui se trouvent dans la partie en haut à droite et la partie en bas à gauche puisque on cherche l'égalité de signes entre les gamma et les corrélations simples. Ce graphique vérifie notre conclusion par rapport aux coefficients. Le fait de trouver des gamma positives liées aux proportions de vote pour quelque candidat comme LASSALLE et POUTOU or logiquement plus la proportion de vote augmente plus la proportion de Blancs nuls diminue mais ce n'est pas le cas dans notre modèle.

Graphique 18 : Comparaison entre les coef de PCR et les corrélations simple



- **Quatrième modèle :** modèle Modélisation de Abstentions_2 l'aide de tous les prédicteurs de premier tour

Dans le cadre de choix de composantes principales qui vont être les niveaux prédicteurs on a choisi les composantes significatives après la régression de la variables Abstentions_2 sur toutes les composantes principales de L'ACP sur les variables de tour 1. on a fixé un seuil de 2.5% au delà de ce valeur les composantes principales ne sont pas significatives.

Tableau 21: RMSE er R2 de la régression sur composantes principales

	RMSE	**R2**
Régression sur toutes les variables	0.0001443698	0.9987
Régression sur toutes les composantes	0.0001459144	0.9987
Régression sur Z1, Z4, Z10 et Z14	0.0002199453	0.9689

Le tableau ci-dessus présente encore une fois la même conclusion avec un perd négligeable en RMSE : 0.0001443698 vs 0.0001459144

En passant par le PCR, le R2 résultant après la régression de Abstentions_2 sur les 4 composantes sélectionnées perd en termes de pouvoir explicative a cause de la perte en termes de quantité d'information dans le modèle. Il passe de 0.9987 avec toutes les variables à 0.9689, avec seulement 4 composantes principales. Un perd plus élevé en RMSE est plus élevé : 0.0001459144 vs 0.0002199453.

Tableau 22 : Coefficients des prédicteurs

	gamm_PCR	corr simple
Abstentions_1	0.2994412	-0.19***
ARTHAUD	-0.3201028	-0.44**
ROUSSEL	-1.2178587	-0.23**
MACRON	-0.5392662	0.082**
LASSALLE	-0.3918030	-0.42**
LE_PEN	-0.6508984	0.64**
ZEMMOUR	-0.2236321	0.02**
MELENCHON	-0.3463709	0.24
HIDALGO	-0.5933991	-0.22**
JADOT	-0.6718163	0.23**
PECRESSE	-0.2440543	-0.21**
POUTOU	-0.6871766	-0.44**
DUPONT_AIGNAN	-0.7092357	-0.45**
Blancs_Nuls_1	-0.5804400	0.32**

Le tableau nous permet d'extraire les variables qui agissent le plus en valeur absolue sur Abstentions_2 sont ROUSSEL, DUPONT_AIGNAN ,LASSALLE ,HIDALGO, JADOT, Blancs_Nuls_1 et POUTOU. L'évaluation de ces estimateurs va être basée sur une comparaison entre les signes des coefficients de corrélation simple et chaque variable et Abstentions_2 . On constate qu'il y a peu de coef gamma qui ont subi le changement de signe : MACRON, ZEMMOUR, PECRESSE. Même souci que le modele precedent puisque on a trouvé des signes gamma non logiques liées aux proportions de vote pour quelque candidats comme LASSALLE, POUTOU et ROUSSEL or logiquement plus la proportion de vote augmente plus la proportion de Abstentions_1 diminue mais ce n'est pas le cas dans notre modèle.

Cependant, il est nécessaire de noter que la stratégie de sélection de composantes principales significatives n'est pas toujours la bonne et que des erreurs peuvent être commises.

3.3 PLS

Nous essayons à présent la régression PLS dans le but d'améliorer la qualité de prédiction de nos modèles dans le cadre de la multicolinéarité.

- **Premier modèle:** Modelisation de MACRON2 l'aide de tous les prédicteurs PLS

La section "TRAINING : % variance explained" montre quelle part de la variance de la variable réponse (MACRON_2) et de la matrice prédicteur (matrice) est expliquée par chaque composante. La section "VALIDATION : RMSEP" montre la racine de l'erreur quadratique moyenne de prédiction (RMSEP) pour chaque composant, telle qu'évaluée par validation croisée "leave-one-out". La RMSEP est une mesure de l'erreur de prédiction du modèle.

Pour la variable de réponse, la première composante explique 64,06% de la variance et la deuxième composante explique 30,34% supplémentaires de la variance, pour un total de 97,73% de variance expliquée par les trois composantes combinées. Ces résultats indiquent que le modèle est capable de capturer une grande partie de la variance à la fois dans la matrice des prédicteurs et dans la variable de réponse en utilisant seulement trois composantes. Cela suggère que le modèle est bien adapté aux données et qu'il est susceptible d'avoir une bonne performance prédictive.

Tableau 23: Sortie R2 du modèle PLS

	1 comps	**2 comps**	**3 comps**
X	60.19	85.07	94.16
MACRON_2	64.06	94.40	97.73

Dans ce cas, le modèle a une RMSEP de 0,04354 pour la première composante, ce qui indique que le modèle a une erreur de prédiction relativement faible**.**

Tableau 24: Sortie RMSE du modèle PLS

	(Intercept)	**1 comps**	**2 comps**	**3 comps**
CV	0.06828	0.04353	0.01748	0.01772
adjCV	0.06828	0.04354	0.01748	0.01171

La sortie de la fonction coef() montre les coefficients des prédicteurs du modèle. Nous pouvons voir que la variable MACRON a un coefficient positif, ce qui suggère une association positive avec MACRON_2. D'autre part, LE_PEN a un coefficient négatif, ce qui indique une association négative avec MACRON_2.

Tableau 25 : Coefficients des prédicteurs

	MACRON_2
Abstentions_1	-0.3967977921
Blancs_Nuls_1	-0.0162298718
ARTHAUD	0.0002218675
ROUSSEL	-0.0154462192
MACRON	0.6961620445
LASSALLE	-0.1057989904
LE_PEN	-0.6811176390
ZEMMOUR	-0.0767484378
MELENCHON	0.3177010890
HIDALGO	0.0217714980
JADOT	0.1579145477

PECRESSE	0.1160843151
POUTOU	-0.0005139096
DUPONT_AIGNAN	-0.0172025019

- **Deuxième modèle:** Modélisation de LE_PEN2 l'aide de tous les prédicteurs PLS

Pour la variable de réponse, la première composante explique 66.40% de la variance de la variable réponse. Pour la deuxième composante, la matrice du prédicteur explique 96.28% de la variance de la variable de réponse. Cela suggère que l'ajout de la deuxième composante améliore significativement la capacité du modèle à expliquer la variation de la variable de réponse, contrairement à l'ajout de la troisième composante.

Tableau 26: Sortie R2 du modèle PLS

	1 comps	**2 comps**	**3 comps**
X	54.97	85.72	94.25
LE_PEN_2	66.40	96.28	96.89

Pour la première composante, le CV RMSEP est de 0,04050, ce qui signifie que le modèle est capable de prédire la variable de réponse avec une erreur moyenne de 0,04050 en utilisant la validation croisée. De même, pour la deuxième composante, le CV RMSEP est de 0,01438, ce qui est inférieur au RMSEP pour 1 composante. Enfin, pour troisième composante, le CV RMSEP est de 0.01390, ce qui est inférieur au RMSEP pour les composantes 1 et 2. Cela suggère que la performance prédictive du modèle s'améliore avec l'ajout encore un peu plus de la troisième composante.

Tableau 27: Sortie RMSE du modèle PLS

	1 comps	**2 comps**	**3 comps**
CV	0.04050	0.01438	0.01391
adjCV	0.04051	0.01437	0.01390

Dans l'ensemble, le modèle semble être bien ajusté car il explique un pourcentage élevé de la variance de la variable de réponse à l'aide de la matrice des prédicteurs. Par ailleurs, les faibles valeurs de RMSEP indiquent que le modèle a une bonne performance prédictive et peut être utilisé pour prédire la variable de réponse avec une précision raisonnable.

Tableau 28 : Coefficients des prédicteurs

	LE_PEN_2
Abstentions_1	-0.187431525
Blancs_Nuls_1	0.021133311
ARTHAUD	0.007175800
ROUSSEL	0.022366854
MACRON	-0.410053381
LASSALLE	0.073901760
LE_PEN	0.926985968
ZEMMOUR	0.083654422
MELENCHON	-0.378017229
HIDALGO	-0.015713904
JADOT	-0.120536071
PECRESSE	-0.060113664

POUTOU	0.004337852
DUPONT_AIGNAN	0.032309809

- **Troisième modèle:** Modélisation de Abstentions_2 l'aide de tous les prédicteurs PLS

Pour le modèle utilisant une seule composante, la valeur du R-carré est de 95,80%, ce qui signifie que le modèle explique environ deux tiers de la variabilité de la variable de résultat. Pour le modèle utilisant deux composantes, la valeur du R-carré est légèrement plus élevée, à 97,52 %, ce qui signifie que le modèle explique une plus grande partie de la variabilité de la variable de résultat.

Tableau 29: Sortie R2 du modèle PLS

	1 comps	**2 comps**
X	66.64	81.31
Abstentions_2	95.80	97.52

La valeur d'interception de 0,07961 dans les deux colonnes représente la valeur RMSEP moyenne obtenue en prédisant la variable de réponse en utilisant uniquement le terme d'interception du modèle.

Les valeurs RMSEP dans les colonnes "1 comps" et "2 comps" représentent les valeurs RMSEP obtenues en incluant une et deux composantes dans le modèle, respectivement. Ces valeurs suggèrent que le modèle à deux composantes a une valeur RMSEP inférieure, ce qui signifie qu'il est le modèle le plus performant.

Les valeurs adjCV sont légèrement inférieures aux valeurs CV correspondantes, ce qui suggère que le modèle peut être surajusté aux données.

Dans l'ensemble, les résultats suggèrent qu'un modèle PLSR à deux composantes est le meilleur modèle pour prédire la variable "Abstentions_2" .

Tableau 30: Sortie RMSE du modèle PLS

	(Intercept)	**1 comps**	**2 comps**
CV	0.07961	0.01753	0.01423
adjCV	0.07961	0.01753	0.01422

Tableau 31 : Coefficients des prédicteurs

	Abstentions_2
Abstentions_1	0.699081258
Blancs_Nuls_1	-0.018397602
ARTHAUD	-0.009749749
ROUSSEL	-0.030025519
MACRON	-0.241184774
LASSALLE	-0.058711990
LE_PEN	-0.336499458
ZEMMOUR	0.013745849
MELENCHON	0.110814869
HIDALGO	-0.025074935
JADOT	-0.023616897
PECRESSE	-0.053166621
POUTOU	-0.009101346

DUPONT_AIGNAN	-0.018113085

- **Quatrième modèle:** Modélisation de Blancs_Nuls_2 l'aide de tous les prédicteurs PLS

Les valeurs suggèrent que le modèle à huit composantes explique un pourcentage plus élevé de la variance de la variable de réponse par rapport au modèle à une composante. On peut voir que le modèle atteint un pourcentage de variance expliquée très élevé (94,57%) avec 8 composantes, ce qui indique que ce modèle est capable de capturer la plupart de la variance dans les données.
En résumé, l'utilisation d'au moins 8 composantes permet d'obtenir une modélisation précise de la relation entre les variables indépendantes et dépendantes.

Tableau 32: Sortie R2 du modèle PLS

	1 comps	**2 comps**	**3 comps**	**4 comps**	**5 comps**	**6 comps**	**7 comps**	**8 comps**
X	66.51	72.87	80.51	96.67	98.55	98.71	99.50	99.82
Blancs_Nuls_2	50.11	75.95	83.54	86.11	89.63	93.71	94.01	94.57

La valeur d'interception de 0,01734 dans les deux colonnes représente la valeur RMSEP moyenne obtenue en prédisant la variable de réponse en utilisant uniquement le terme d'interception du modèle.

Les valeurs RMSEP entre la composante 1 et 8 diminuent de moitié. Ces valeurs suggèrent que le modèle à 8 composantes a la valeur RMSEP la plus inférieure, ce qui signifie qu'il est le modèle le plus performant. Les valeurs adjCV sont légèrement inférieures aux valeurs CV correspondantes, ce qui suggère que le modèle peut être surajusté aux données.

Tableau 33: Sortie RMSE du modèle PLS

	1 comps	**2 comps**	**3 comps**	**4 comps**	**5 comps**	**6 comps**	**7 comps**	**8 comps**
CV	0.01249	0.009608	0.007765	0.007538	0.007166	0.007891	0.006327	0.006400
adjCV	0.01249	0.009594	0.007737	0.007543	0.007157	0.007863	0.006318	0.006386

Tableau 34 : Coefficients des prédicteurs

	Blancs_Nuls_2
Abstentions_1	-1.482490e-01
Blancs_Nuls_1	3.349706e-02
ARTHAUD	-9.989982e-05
ROUSSEL	6.388606e-02
MACRON	-1.944614e-01
LASSALLE	2.643786e-01
LE_PEN	-2.719770e-02
ZEMMOUR	-5.487103e-02
MELENCHON	6.157204e-02
HIDALGO	5.459961e-02
JADOT	-2.475905e-02
PECRESSE	-3.427571e-02
POUTOU	1.206312e-02
DUPONT_AIGNAN	-6.082694e-03

3.4 Comparaison des modèles

Tableau 35: Comparaison des coefficients des prédicteurs et des corrélations des modèles

	MCO				RCP			
Variable	**MACRON**	**LE PEN**	**BLANCS_NULS**	**ABSTENTIONS**	**MACRON**	**LE PEN**	**BLANCS_NULS**	**ABSTENTIONS**
Abstentions_1	0.03150187	0.09452174	-0.0135390230	0.88566557	0.88869841	1.8592356	-0.07554219	0.2994412
Blancs_Nuls_1	-0.98634251	0.20987149	1.2211117455	0.59608742	-0.64491176	1.4231359	0.31474159	-0.5804400
ARTHAUD	4.96728895	0.52911827	-1.6959425070	-2.78579582	3.28505396	0.7902248	0.00362356	-0.3201028
ROUSSEL	-0.10208363	-0.36309998	0.8332651109	0.62655860	-0.80261621	1.1712853	0.80443349	-1.2178587
MACRON	1.29569703	-0.22117603	-0.0503219722	-0.02356905	-0.01128135	2.2106592	-0.12248514	-0.5392662
LASSALLE	0.41987973	0.22329947	0.2272967688	0.12785250	0.85160509	2.3305518	0.37726146	-0.3918030
LE_PEN	-0.04101779	1.16944757	-0.0005998119	-0.12994493	-0.12452710	2.8296652	0.05762821	-0.6508984
ZEMMOUR	0.16412410	0.62790860	-0.0731547965	0.28587720	1.07008830	0.5426811	-0.06833560	-0.2236321
MELENCHON	0.54189470	0.04482591	0.0760231076	0.33929836	0.38336681	2.1065524	0.04793250	-0.3463709
HIDALGO	-0.01494338	0.84860309	0.8537413693	-0.70298927	-1.94520632	2.1135011	0.42976884	-0.5933991
JADOT	1.38056962	-0.35507648	0.0396379409	-0.07737149	0.08423118	1.4751500	0.50486436	-0.6718163
PECRESSE	0.55248945	0.37182509	0.0233142264	0.04912708	-0.50087328	0.6008762	-0.04413244	-0.2440543
POUTOU	-3.44222900	4.32216340	2.0472883358	-1.92662176	-2.88413294	2.3058307	0.53938866	-0.6871766
DUPONT_AIGNAN	-1.84142979	1.04593368	0.7915672526	1.02094227	-2.89162065	1.6286894	0.54530423	-0.7092357
	PLS				**Corr(X,Y)**			
	MACRON	LE PEN	BLANCS_NULS	ABSTENTIONS	MACRON	LE PEN	BLANCS_NULS	ABSTENTIONS
Abstentions_1	-0.3967977921	-0.187431525	-1.482490e-01	0.699081258	-0.22***	-0.14***	-0.16**	-0.19***
Blancs_Nuls_1	-0.0162298718	0.021133311	3.349706e-02	-0.018397602	-0.08	0.41***	0,61**	0.32**
ARTHAUD	0.0002218675	0.007175800	-9.989982e-05	-0.009749749	0.1	0.32***	0.16	-0.44**
ROUSSEL	-0.0154462192	0.022366854	6.388606e-02	-0.030025519	0.38***	0.37***	0.78***	-0.23**
MACRON	0.6961620445	-0.410053381	-1.944614e-01	-0.241184774	0.94***	-0.17	0.23*	0.082**
LASSALLE	-0.1057989904	0.073901760	2.643786e-01	-0.058711990	0.1	0.29**	0.76***	-0.42**
LE_PEN	-0.6811176390	0.926985968	-2.719770e-02	-0.336499458	-0.05	0.92**	0.36***	0.64**
ZEMMOUR	-0.0767484378	0.083654422	-5.487103e-02	0.013745849	0.23*	0.36***	0.16	0.02**
MELENCHON	0.3177010890	-0.378017229	6.157204e-02	0.110814869	0.29**	-0.27**	0.29**	0.24
HIDALGO	0.0217714980	-0.015713904	5.459961e-02	-0.025074935	0.47***	0.04	0.78***	-0.22**
JADOT	0.1579145477	-0.120536071	-2.475905e-02	-0.023616897	0.86***	-0.28**	0.27**	0.23**
PECRESSE	0.1160843151	-0.060113664	-3.427571e-02	-0.053166621	0.46***	-0.11	0.05	-0.21**
POUTOU	-0.0005139096	0.004337852	1.206312e-02	-0.009101346	0.37***	0.30**	0.67***	-0.44**
DUPONT_AIGNAN	-0.0172025019	0.032309809	-6.082694e-03	-0.018113085	0.20*	0.48***	0.21*	-0.45**

Tableau 35: Comparaison des RMSE et R2 des modèles

	RMSE				R squared			
Approche	**MACRON**	**LE PEN**	**BLANCS_NULS**	**ABSTENTIONS**	**MACRON**	**LE PEN**	**BLANCS_NULS**	**ABSTENTIONS**
MCO	0.000069357	0.000140781	0.0000201483	0.000144369	0.9997	0.999	0.997	0.9987
RCP	8.954271e-05	0.001958971	2.626683e-05	0.0002199453	0.984	0.5811	0.9289	0.9689
PLS	0.01772	0.01391	0.006386	0.01422	97.73	96.89	94.57	97.52

Nous pouvons constater en comparant les coefficients des prédicteurs et des corrélations des modèles que les coefficients sont différents pour chaque modèle.Les différences entre les coefficients de régression obtenus à partir de différentes méthodes d'analyse de données peuvent être dues aux différences dans les algorithmes utilisés pour ajuster les modèles aux données. En effet, les méthodes MCO, PCR et PLS utilisent des approches différentes pour estimer les coefficients de régression. Le MCO utilise une approche classique qui minimise la somme des carrés des résidus entre les valeurs prédites et les valeurs réelles. PCR utilise une approche basée sur une réduction de la dimension de l'espace des variables explicatives par une décomposition en valeurs singulières. PLS utilise une approche basée sur une régression partielle et une maximisation de la covariance entre les variables explicatives et la variable à expliquer.

Ces différences dans les algorithmes peuvent entraîner des différences dans les coefficients de régression obtenus, en particulier si les variables explicatives sont fortement corrélées ce qui est le cas pour nos 14 variables explicatives.

Il est donc important de comparer les résultats de différents modèles afin de mesurer leur performance. En ce qui concerne les mesures de performance, il est important de prendre en compte plusieurs critères tels que le RMSE et le R-squared. Pour comparer les résultats de RMSE et de R-squared des différents modèles, il convient de prendre en compte plusieurs critères. Tout d'abord, le RMSE (Root Mean Square Error) est une mesure de la différence entre les valeurs prédites par un modèle et les valeurs réelles. Plus le RMSE est faible, meilleure est la précision du modèle. En général, il est recommandé de choisir le modèle avec le RMSE le plus faible.

Ensuite, le R-squared est une mesure de la proportion de la variance des données expliquée par le modèle. Plus le R-squared est proche de 1, meilleure est l'ajustement du modèle. Cependant, il convient de noter que le R-squared ne prend pas en compte la complexité du modèle et peut donc conduire à la sur-estimation de la qualité du modèle.

En utilisant ces critères, on peut conclure que le modèle PLS est le moins précis en termes de RMSE, et présente les moins bons ajustements en termes de R-squared. Le modèle MCO présente les meilleures performances en termes de RMSE pour toutes les variables, et avec des R-squared supérieurs pour toutes les variables. Le modèle RCP présente également de bonnes performances en termes de RMSE, mais avec un R-squared relativement faible pour la variable LE PEN. **En conclusion, le modèle MCO est celui qui semble être le meilleur compromis en termes de RMSE et de R-squared, avec des performances solides pour toutes les variables.**

5. Modélisation multidimensionnelle PLS

Pour la variable de réponse "MACRON_2", la première composante explique 29,61% de la variance, tandis que les quatre premières composantes expliquent ensemble 98,02% de la variance. Pour la variable de réponse "LE_PEN_2", la première composante explique 19,73% de la variance etles quatre premières composantes

expliquent ensemble 96,78% de la variance. Pour la variable de réponse "Blancs_Nuls_2", les composantes 3 et 4 expliquent ensemble 83,07% de la variance. Pour la variable de réponse "Abstentions_2", les quatre premières composantes expliquent ensemble 97,8% de la variance.

Tableau 37: Sortie R2 du modèle PLS

	1 comps	2 comps	3 comps	4 comps
X	66.66	85.59	94.25	97.26
MACRON_2	29.61	90.89	97.54	98.02
LE_PEN_2	19.73	96.60	96.76	96.78
Blancs_Nuls_2	45.57	45.78	54.36	83.07
Abstentions_2	95.69	96.06	97.63	97.84

Les résultats suggèrent que les deux premières composantes ont une contribution significative à la régression pour toutes les variables de réponse, car l'erreur RMSEP diminue considérablement lorsqu'on passe de 0 à 1 ou 2 composantes. Cependant, l'ajout de la troisième et de la quatrième composante n'a pas autant d'effet sur l'erreur de prédiction. Ces résultats suggèrent que la régression PLS multiple est capable de modéliser les relations linéaires entre les variables explicatives et la variable de réponse avec une précision raisonnable.

Tableau 38: Sortie RMSE du modèle PLS

MACRON_2	(Intercept)	1 comps	2 comps	3 comps	4 comps
CV	0.06828	0.05858	0.02193	0.01208	0.01112
adjCV	0.06828	0.05858	0.02192	0.01207	0.01111

Tableau 39: Sortie RMSE du modèle PLS

LE_PEN_2	(Intercept)	1 comps	2 comps	3 comps	4 comps
CV	0.06563	0.06023	0.01378	0.01377	0.01388
adjCV	0.06563	0.06022	0.01376	0.01376	0.01387

Tableau 40: Sortie RMSE du modèle PLS

Blancs_Nuls_2	(Intercept)	1 comps	2 comps	3 comps	4 comps
CV	0.01734	0.01291	0.01304	0.01240	0.007986

adjCV	0.01734	0.01291	0.01304	0.01239	0.007981

Tableau 41: Sortie RMSE du modèle PLS

Abstentions_2	(Intercept)	1 comps	2 comps	3 comps	4 comps
CV	0.07961	0.01781	0.01722	0.01399	0.01346
adjCV	0.07961	0.01779	0.01721	0.01398	0.01345

Tableau 42: Coefficients des prédicteurs

	MACRON_2	LE_PEN_2	Blancs_Nuls_2	Abstentions_2
Abstentions_1	-0.3516418	-0.1870590	-0.1802969	0.7179747
Blancs_Nuls_1	-0.0432101	0.0206420	0.0514318	-0.0290257
ARTHAUD	0.0023001	0.0040054	0.0043876	-0.0107353
ROUSSEL	-0.0624264	0.0257051	0.0835801	-0.0470753
MACRON	0.9732957	-0.4972260	-0.1805894	-0.2959265
LASSALLE	-0.3398273	0.0816581	0.3847755	-0.1277775
LE_PEN	-0.5694285	0.9311368	-0.1029458	-0.2575620
ZEMMOUR	-0.0500022	0.1081207	-0.1046838	0.0473655
MELENCHON	0.1604110	-0.2863028	-0.0431357	0.1707171
HIDALGO	-0.0217190	-0.0187420	0.0829016	-0.0427446
JADOT	0.1912184	-0.1249192	-0.0366759	-0.0295420
PECRESSE	0.1400203	-0.0880559	0.0143074	-0.0665934
POUTOU	-0.0108629	0.0029593	0.0201741	-0.0123405
DUPONT_AIGNAN	-0.0181273	0.0280777	0.0067694	-0.0167346

Conclusion

En conclusion, l'analyse des élections et des résultats électoraux est une tâche complexe qui nécessite l'utilisation de différentes techniques d'analyse de données et de modèles.

Dans ce rapport, nous avons présenté une analyse approfondie des données et des modèles utilisés pour prédire l'issue d'une élection entre différents candidats.

Nous avons utilisé l'analyse exploratoire des données, l'analyse de composantes principales et plusieurs modèles de régression unidimensionnels et multidimensionnels pour prédire les résultats électoraux.

Nous avons comparé les résultats de différents modèles en utilisant des critères tels que le RMSE et le R-squared, et avons conclu que le modèle MCO est celui qui semble être le meilleur compromis en termes de RMSE et de R-squared, avec des performances solides pour toutes les variables.

Nous avons également souligné l'importance de comparer les résultats de différents modèles afin de mesurer leur performance.

En fin de compte, cette analyse des élections et des résultats électoraux fournit une base solide pour les partis politiques, les médias et les électeurs pour comprendre les tendances électorales et prédire les résultats des élections futures.

Annexes

Annexe 01: statistiques descriptives

	Abstentions_1	ARTHAUD	ROUSSEL	MACRON	LASSALLE	LE_PEN	ZEMMOUR	MELENCHON	HIDALGO
Min.	0.1919	0.0009055	0.00179	0.04953	0.002967	0.01837	0.005228	0.03887	0.00345
1st Qu.	0.2178	0.0035626	0.0141	0.17214	0.017538	0.15836	0.041196	0.12441	0.009771
Median	0.2373	0.0046695	0.01686	0.19218	0.023638	0.18446	0.049659	0.14623	0.013148
Mean	0.2716	0.004406	0.01678	0.1898	0.027895	0.17539	0.048838	0.14777	0.01305
3rd Qu.	0.2583	0.0052888	0.0207	0.21499	0.032593	0.20365	0.055415	0.16585	0.015486
Max.	0.6913	0.0102855	0.034	0.28817	0.092793	0.27961	0.100449	0.33584	0.027543
	JADOT	PECRESSE	POUTOU	DUPONT_AIGNAN	Abstentions_2	MACRON_2	LE_PEN_2	Blancs_Nuls_1	Blancs_Nuls_2
Min.	0.0032	0.00968	0.002190	0.004927	0.2104	0.1273	0.0517	0.003719	0.01315
1st Qu.	0.02385	0.02989	0.005141	0.013621	0.2404	0.3290	0.2623	0.015498	0.05751
Median	0.02902	0.03553	0.005955	0.016212	0.2519	0.3641	0.3032	0.017013	0.06512
Mean	0.0297	0.03573	0.005816	0.015804	0.2836	0.3599	0.2902	0.017366	0.06622
3rd Qu.	0.03627	0.03970	0.006620	0.017935	0.2665	0.4073	0.3249	0.019442	0.07669
Max.	0.05867	0.14211	0.009911	0.025849	0.6518	0.5911	0.4097	0.023622	0.10688

Annexe 02; Matrice de corrélation

	Abstentions_1	**ARTHAUD**	**ROUSSEL**	**MACRON**	**LASSALLE**	**LE_PEN**	**ZEMMOUR**	**MELENCHON**	**HIDALGO**	**JADOT**	**PECRESSE**	**POUTOU**	**DUPONT_AIGNAN**	**Abstentions_2**	**MACRON_2**	**LE_PEN_2**	**Blancs_Nuls_1**	**Blancs_Nuls_2**
Abstentions_1	1.0000000	0.1358473	-0.3301850	-0.24529744	-0.09064738	-0.28995242	-0.3182380827	-0.01262790	-0.04858467	-0.24355211	0.11667047	0.1513360350	0.07611247	-0.19700745	-0.22332899	-0.14621211	-0.26665873	-0.16069472
ARTHAUD	0.13584731	1.0000000	0.1919887	0.25596936	-0.11090838	0.36354728	-0.4205570155	-0.22182600	0.15189956	-0.12733989	0.41277600	0.4547524470	0.48468358	-0.44041443	0.10368634	0.32147735	0.34405276	0.15716057
ROUSSEL	-0.33018503	0.1919887	1.0000000	0.38053372	0.58666518	0.49287031	0.2944477914	0.09498880	0.59685655	0.26978600	0.15559229	0.5521151834	0.14684954	-0.23343840	0.38257661	0.36753637	0.41892510	0.78114685
MACRON	-0.24529744	0.2559694	0.3805337	1.00000000	0.06905497	0.10887325	0.2133337342	0.01730153	0.41747221	0.77918349	0.47219895	0.4043765049	0.34256739	0.08211800	0.93786974	-0.16765795	-0.02397892	0.22602728
LASSALLE	-0.09064738	-0.1109084	0.5866652	0.06905497	1.00000000	0.24665743	0.2836105730	-0.05717086	0.65319083	0.06393998	0.10786564	0.4514021656	0.10724180	-0.41300006	0.09934415	0.28969333	0.34528700	0.75715636
LE_PEN	-0.28995242	0.3635473	0.4928703	0.10887325	0.24665743	1.00000000	0.4200360958	-0.30058378	0.08662305	-0.06077358	-0.04447175	0.3911185109	0.51672921	-0.36624824	-0.05157303	0.92001001	0.36565307	0.35863225
ZEMMOUR	-0.31823808	-0.4205570	0.2944478	0.21333373	0.28361057	0.42003610	1.0000000000	-0.03322841	0.01187998	0.33528945	0.07928735	-0.0009238794	0.29709579	0.02998198	0.23312622	0.35902427	-0.11071124	0.16218863
MELENCHON	-0.01262790	-0.2218260	0.0949888	0.01730153	-0.05717086	-0.30058378	-0.0332284082	1.00000000	0.23013875	0.33272127	-0.19327564	0.0547769547	-0.26379660	0.24103889	0.28811321	-0.27050599	0.03262177	0.29258054
HIDALGO	-0.04858467	0.1518996	0.5968566	0.41747221	0.65319083	0.08662305	0.0118799805	0.23013875	1.00000000	0.39501407	0.15121951	0.5458087245	0.12455214	-0.22781448	0.46714087	0.03687213	0.29515567	0.77550930
JADOT	-0.24355211	-0.1273399	0.2697860	0.77918349	0.06393998	-0.06077358	0.3352894489	0.33272127	0.39501407	1.00000000	0.17454436	0.3707956191	0.19610916	0.23477519	0.85759230	-0.28265339	-0.06580301	0.27460822
PECRESSE	0.11667047	0.4127760	0.1555923	0.47219895	0.10786564	-0.04447175	0.0792873465	-0.19327564	0.15121951	0.17454436	1.00000000	0.1295013883	0.37402154	-0.21074681	0.45725844	-0.10591796	-0.07658129	0.05412289
POUTOU	0.15133604	0.4547524	0.5521152	0.40437650	0.45140217	0.39111851	-0.0009238794	0.05477695	0.54580872	0.37079562	0.12950139	1.0000000000	0.40199571	-0.44815174	0.36538156	0.30256244	0.36702693	0.67136128
DUPONT_AIGNAN	0.07611247	0.4846836	0.1468495	0.34256739	0.10724180	0.51672921	0.2970957858	-0.26379660	0.12455214	0.19610916	0.37402154	0.4019957149	1.00000000	-0.44067297	0.19641648	0.47728964	0.09280580	0.21216905
Absten	-0.197	-0.440	-0.233	0.0821	-0.413	-0.366	0.0299	0.2410	-0.227	0.2347	-0.210	-0.448	-0.440	1.0000	0.1929	-0.472	-0.513	-0.460

tions_2	00745	4144	4384	1800	00006	24824	819831	3889	81448	7519	74681	1517417	67297	0000	5505	25762	14907	73998
MACRON_2	-0.22332899	0.1036863	0.3825766	0.93786974	0.09934415	-0.05157303	0.2331262181	0.28811321	0.46714087	0.85759230	0.45725844	0.3653815618	0.19641648	0.19295505	1.00000000	-0.31876897	-0.08272789	0.27521091
LE_PEN_2	-0.14621211	0.3214773	0.3675364	-0.16765795	0.28969333	0.92001001	0.3590242665	-0.27050599	0.03687213	-0.28265339	-0.10591796	0.3025624353	0.47728964	-0.47225762	-0.31876897	1.00000000	0.41120476	0.35774797
Blancs_Nuls_1	-0.26665873	0.3440528	0.4189251	-0.02397892	0.34528700	0.36565307	-0.1107112387	0.03262177	0.29515567	-0.06580301	-0.07658129	0.3670269265	0.09280580	-0.51314907	-0.08272789	0.41120476	1.00000000	0.61113004
Blancs_Nuls_2	-0.16069472	0.1571606	0.7811468	0.22602728	0.75715636	0.35863225	0.1621886326	0.29258054	0.77550930	0.27460822	0.05412289	0.6713612811	0.21216905	-0.46073998	0.27521091	0.35774797	0.61113004	1.00000000