2016

Les ressorts socio-économiques de l'élection présidentielle américaine

Aux Etats-Unis, la course à la présidentielle 2016 pour succéder à Barack Obama à la Maison-Blanche a été remportée au soir du 8 novembre par le républicain Donald Trump. Lors du scrutin, le milliardaire populiste s'est emparé de plusieurs Etats-clés face à sa rivale, la démocrate Hillary Clinton, en dépit des multiples sondages prédisant l'extrême inverse.

L'objectif est ici de comprendre les raisons socio-économiques d'un tel vote grâce à diverses méthodes de Datamining

Membres du groupe :

Tarmoul lydia
Koskakova galina
Larbani wanis
Bouchebbah arslane
Caldichoury arnaud

Chargé de cours : Chalmel Vincent

PROMO : 2016/2017.

Sommaire

Introduction

Les États-Unis sont une république fédérale. « Fédérale » signifie qu'elle est composée de 50 États et de 14 territoires autonomes qui se rassemblent pour former une Nation.

L'élection présidentielle américaine se déroule dans chacun de ces États. Les résultats des votes sont récoltés par comté, chaque comté appartenant à un Etat donné. Le scrutin est universel indirect.

Dans chaque État, contrairement aux élections présidentielles en France, les citoyens américains ne votent pas directement pour le candidat à la présidence mais désignent les grands électeurs de leur État qui éliront ensuite le président. Leur nombre varie de 3 au minimum dans les plus petits États (comme le District Of Columbia qui abrite la capitale Washington) à 55 pour la Californie, le plus peuplé des États du pays. Il y a au total 538 Grands Électeurs.

Le candidat qui remporte la majorité des voix d'un État se voit attribuer tous les Grands électeurs de celui-ci. Seuls le Nebraska et le Maine appliquent la proportionnelle. Ce système peut aboutir à une situation où le Président élu n'a pas obtenu la majorité des suffrages de la population. C'est ce qui s'est passé en 2000 lors de la première élection de Georges W. Bush. Le candidat qui obtient 270 Grands électeurs ou plus devient président des États-Unis.

Ce fort nombre de comtés engendre donc une forte variabilité des votes au niveau national (en fonction de la localisation du comté et de ses caractéristiques socio-économiques).

En plus de cela, la variabilité peut être forte au sein même de chaque comté. En effet, même si traditionnellement certains États ont tendance à voter toujours en faveur des républicains (le Texas par exemple) ou des démocrates (la Californie ou le Maine, entre autres), une douzaine d'États ("swing states") peuvent basculer d'un côté ou de l'autre et ainsi déterminer le résultat de l'élection présidentielle. C'est justement dans ces États qu'Hillary Clinton a sous-performé, laissant le champ libre à son rival.

On cherchera par conséquent ici à expliquer les ressorts sociaux et économiques de l'élection présidentielle américaine de 2016. L'objectif est en effet ici de parvenir :

- **à diviser l'ensemble des comtés américains en sous-populations homogènes**

- **à construire un modèle explicatif des résultats de l'élection de 2016 dans chacune des classes de comtés créées, et ce à partir de données sur les résultats présidentiels de 2012 et de données socio-économiques sur chaque comté.**

1) Etude exploratoire

a. Présentation des données

La base que nous avons à disposition est une base qui présente les votes de 2012 et 2016 aux élections présidentielles par comté.

Un individu (ie une ligne dans la base de données) correspond donc à un comté américain. La population totale s'élève donc à 3112 comtés. La base de données comprend également 81 variables ou colonnes.

La base est issue du site : https://www.kaggle.com/joelwilson/2012-2016-presidential-elections. On dispose également du dictionnaire complet des variables.

b. Statistiques univariées & choix des variables

Les variables présentes dans la base de données sont quantitatives ou explicatives. Il y a plus précisement 4 variables qualitatives :

- State_abreviation
- County_name
- State_abbrv
- Area_name

Nous avons estimé que ces variables n'apportaient pas de valeur ajoutée pour notre classification et nos régressions futures. Elles sont donc écartées dès le départ.

D'autres variables sont redondantes ou entièrement corrélées aux résultats de l'élection présidentielle de 2016 et doivent donc être écartées. En voici la liste:
['X','combined_fips','votes_dem_2016','votes_gop_2016','total_votes_2016','Clinton','Trump', 'diff_2016','per_point_diff_2016','state_abbr','county_name', 'FIPS','total_votes_2012', 'votes_dem_2012', 'votes_gop_2012', 'county_fips','state_fips', 'Obama', 'Romney','diff_2012', 'state_abbreviation','fips', 'area_name', 'population2014','population2010','POP010210', 'Clinton_Obama', 'Trump_Romney', 'Trump_Prediction','Clinton_Prediction', 'Trump_Deviation', 'Clinton_Deviation']

Une fois les variables inintéressantes supprimées, il est primordial de réaliser l'étude exploratoire. L'objectif est de décrire les différentes variables grâce à des indicateurs statistiques simples. C'est également lors de cette étude que l'on réalise :

- la gestion des valeurs manquantes (aucune dans notre cas)
- l'identification des valeurs / individus aberrants

Nous avons obtenu les statistiques descriptives suivantes à l'aide d'un Proc Means :

Le Système SAS

Procédure MEANS

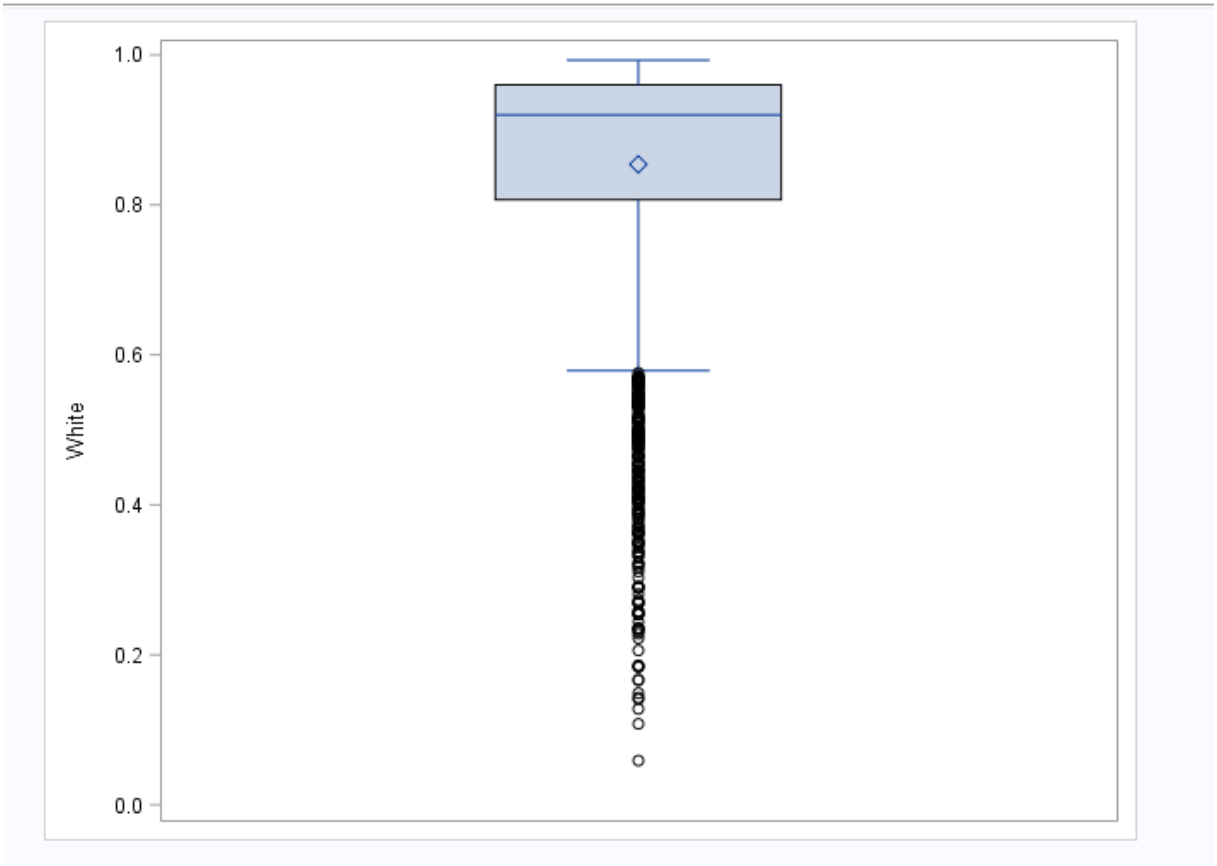
Variable	Libellé	N	Moyenne	Ecart-type	Minimum	Maximum
A	A	3112	1555.50	898.5013448	0	3111.00
per_point_diff_2012	per_point_diff_2012	3112	-0.2128900	0.2954996	-0.9241379	0.8734824
population_change	population_change	3112	0.4501928	4.1955817	-17.0000000	72.9000000
AGE135214	AGE135214	3112	5.8869537	1.1830652	1.5000000	13.3000000
AGE295214	AGE295214	3112	22.5375964	3.3336002	7.4000000	40.5000000
age65plus	age65plus	3112	17.6401992	4.3870730	4.1000000	52.9000000
SEX255214	SEX255214	3112	49.9557519	2.2053030	30.1000000	56.8000000
White	White	3112	0.8540026	0.1578842	0.0590000	0.9930000
Black	Black	3112	0.0930045	0.1448295	0	0.8510000
RHI325214	RHI325214	3112	1.9743252	6.5467441	0	92.2000000
RHI425214	RHI425214	3112	1.3605077	2.5492372	0	42.4000000
RHI525214	RHI525214	3112	0.1027314	0.4071945	0	12.7000000
RHI625214	RHI625214	3112	1.8546594	1.3546582	0	29.4000000
Hispanic	Hispanic	3112	0.0904714	0.1353501	0.0020000	0.9580000
RHI825214	RHI825214	3112	77.3661954	19.6302684	3.1000000	98.6000000
POP715213	POP715213	3112	86.4286632	4.3931203	50.8000000	99.8000000
POP645213	POP645213	3112	4.4927378	5.5208956	0	51.3000000
NonEnglish	NonEnglish	3112	9.1393959	11.4002209	0	95.6000000
Edu_highschool	Edu_highschool	3112	84.5137532	6.9121468	45.0000000	99.0000000
Edu_batchelors	Edu_batchelors	3112	19.7427378	8.8301043	3.2000000	74.4000000
VET605213	VET605213	3112	6809.81	16340.26	2.0000000	331642.00
LFE305213	LFE305213	3112	23.0901671	5.3583618	8.2000000	44.2000000
HSG010214	HSG010214	3112	42946.17	125433.12	50.0000000	3482516.00

Ces premières statistiques nous permettent de faire quelques descriptions de variables à partir de l'information chiffrée.

D'après le tableau, on remarque une certaine hétérogénéité entre les variables. Pour chacune des variables, il y a une grande différence entre sa valeur minimale et maximale.

Par exemple en moyenne, on remarque il y a 85% de citoyens américains blancs (variable White) par comté et 9% de citoyens noirs (variable Black) par comté. Néanmoins, pour la variable Black, l'écart-type est bien plus important par rapport à la moyenne. Cette variable est donc très dispersée, et peut s'avérer être discriminante pour les prochaines classifications/régressions.

Nous pouvons générer des boites à moustache afin d'avoir un aperçu rapide des caractéristiques de chaque variable. Pour la variable « white » par exemple :



Cette boite à moustache nous renseigne donc sur les valeurs minimums et maximums de la proc means de cette variable, ainsi que le 1er et 3e quartile. La différence entre ces valeurs nous donne l'amplitude totale qui est la dispersion de ces valeurs.

On utilise une proc univariate pour mieux illustrer cette boite a moustache et avoir la vraie valeur des quartiles :

Le Système SAS

Procédure UNIVARIATE
Variable : White (White)

Moments

N	3112	**Somme des poids**	3112
Moyenne	0.85400257	**Somme des observations**	2657.656
Ecart-type	0.15788425	**Variance**	0.02492744
Skewness	-1.8776422	**Kurtosis**	3.42883383
Somme des carrés non corrigée	2347.19431	**Somme des carrés corrigée**	77.549252
Coeff Variation	18.4875611	**Std Error Mean**	0.00283021

Mesures statistiques de base

Emplacement		Variabilité	
Moyenne	0.854003	**Ecart-type**	0.15788
Médiane	0.920000	**Variance**	0.02493
Mode	0.970000	**Intervalle**	0.93400
		Ecart interquartile	0.15300

Tests de tendance centrale : Mu0=0

Test	Statistique		P-value	
t de Student	**t**	301.7451	**Pr > |t|**	<.0001
Signe	**M**	1556	**Pr >= |M|**	<.0001
Rang signé	**S**	2421914	**Pr >= |S|**	<.0001

Quantiles (Définition 5)

Niveau		Quantile
100Max	**100%**	0.993
99%		0.985
95%		0.979
90%		0.974
75%	**Q3**	0.960
50%	**Médiane**	0.920
25%	**Q1**	0.807
10%		0.627
5%		0.498
1%		0.273
0%	**Min**	0.059

Observations extrêmes

Le plus bas		Le plus haut	
Valeur	**Obs**	**Valeur**	**Obs**
0.059	2383	0.988	1782
0.108	2392	0.988	1869
0.128	3002	0.988	2707

Observations extrêmes

Le plus bas		Le plus haut	
Valeur	**Obs**	**Valeur**	**Obs**
0.141	1518	0.990	1813
0.143	1497	0.993	1835

c. Statistiques bivariées & étude de la corrélation entre les variables

Sélection des variables avec la corrélation de Spearman

Les statistiques descriptives seules ne suffisent pas. Il est essentiel de réaliser des analyses à l'aide de statistiques bivariées. En particulier, il apparaît primordial d'étudier la corrélation :

- des variables explicatives avec la variable cible
- mais également entre les variables explicatives elles-mêmes

Il existe deux types d'algorithmes pour résoudre cette problématique (paramétrique et non paramétrique). Les algorithmes paramétriques dépendent des hypothèses faites sur les données, c.-à-d., dans le cas de la méthode **Pearson**, il est nécessaire d'avoir certaines informations en amont, comme la distribution, l'échelle et les dépendances des différentes variables. La détermination des lois de chacune des variables est ardue, c'est pourquoi nous avons privilégié la méthode **Spearman** non-paramétrique. Aucune hypothèse de normalité sur les distributions des variables n'est nécessaire avec cette méthode.

Etant donné qu'on dispose déjà d'un dataframe, la première chose à faire est de calculer les corrélations avec la fonction rcorr depuis package Hmisc qui retourne une liste contenant les coefficients de corrélation ainsi qu'une p-value qui représente la significativité. Aussi, il est nécessaire de rappeler en cas d'oubli dans la partie pré-processing, cette fonction supprime les valeurs NA, plutôt que toute la ligne ou la colonne.

correlation_election
.xls

En utilisant la méthode non-paramétrique **Spearman**, il est possible d'avoir une première idée des corrélations entre les variables. Le résultat peut être affiché en heatmap en fonction des p-

values ainsi que les coefficients de corrélation. Malheureusement la qualité du graphique se dégrade avec la croissance du nombre de variables (50 dans notre cas).

Par conséquent, on a choisi de représenter la matrice sous Excel avec comme légende 0 ' ' 0.3 '.' 0.6 ',' 0.8 '+' 0.9 '*' 0.95 'B' 1. Ainsi, par exemple, les coefficients de corrélation entre 0 et 0.3 sont remplacés par un espace.

Cela permet une identification rapide des variables très corrélées (>0.9 ou <-0.9) avec d'autres variables, à savoir. En voici la liste :

- NonEnglish
- Hispanic
- RHI525214 = Native Hawaiian and Other Pacific Islander alone, percent, 2014
- VET605213 = Veterans, 2009-2013
- HSG010214 = Housing units, 2014
- HSD410213 = Households, 2009-2013
- BZA010213 = Private nonfarm establishments, 2013
- BZA110213 = Private nonfarm employment, 2013
- NES010213 = Nonemployer establishments, 2013
- SBO001207 = Total number of firms, 2007

Sélection des variables avec l'indicateur d'Inflation de la Variance

Néanmoins, on se limite ici à la corrélation entre paires de variables, et on ne résout pas complètement le problème de multicolinéarité. En effet, il est possible qu'on obtienne de petites corrélations entre paires de variables alors qu'une forte dépendance linéaire existe entre trois variables ou plus. Ainsi, il est plus judicieux de s'appuyer sur d'autres méthodes comme le **VIF, ou facteur d'inflation de la variance** pour la détection de la multi-colinéarité.

Prenons le cas simple d'un modèle avec des variables corrélées

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i$$

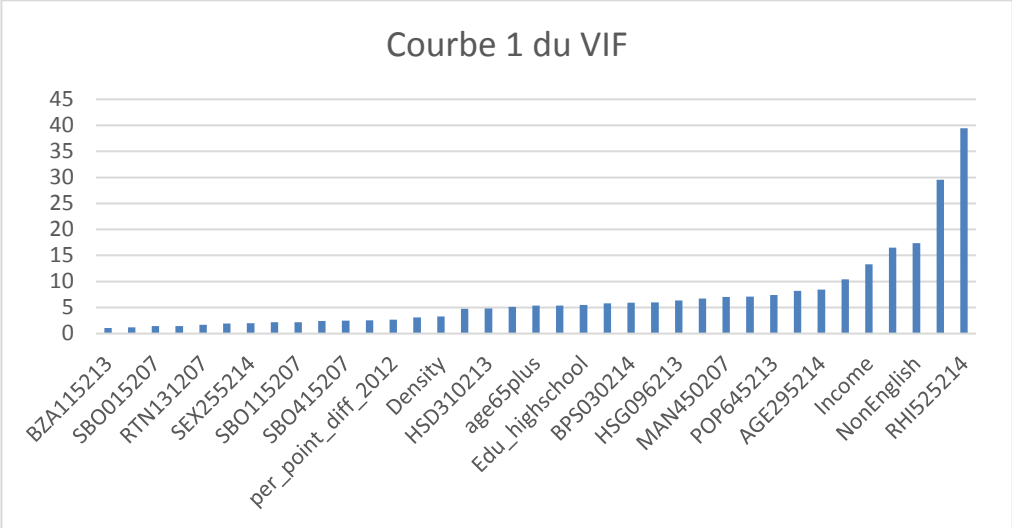
Le VIF associé à chaque variable est exprimé comme suit: $VIF_k = \frac{1}{1 - R_k^2}$, où R^2_k est le résultat de la régression du $k^{ième}$ régresseur sur les régresseurs restants.

La variance du coefficient associé au $k^{ième}$ régresseur est liée au VIF comme le montre la formule suivante : $Var(b_k) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2} \times \frac{1}{1 - R_k^2}$

Ainsi, si une variable explicative est très corrélée avec les autres variables explicatives, alors la variance du coefficient Bk associée à cette variable est enflée, et la variable peut être rejetée lors d'un test de Student.

Ainsi, le VIF mesure de combien la variance d'un coefficient est enflée par l'existence d'une corrélation avec les variables restantes du modèle. En règle générale, un VIF > 15 est le signe d'une forte multi-colinéarité, et il convient donc de s'intéresser de plus près à la variable problématique.

La figure suivante a été obtenue à partir d'une régression multiple sur les résultats 2016 de Clinton.



En appliquant la méthode VIF, on s'intéresse aux variables dépassant le seuil de 15. Cela correspond à des variables très corrélées aux autres variables de la future régression. En voici la liste : ['INC110213','NonEnglish', 'VET605213', 'RHI525214', 'RTN130207', 'BZA110213', 'BZA010213', 'RHI625214', 'NES010213', 'Hispanic', 'HSG010214', 'HSD410213', 'SBO001207', 'RHI425214', 'RHI825214', 'RHI325214', 'Black', 'White']

On examine alors ces variables ciblées les unes après les autres et on questionne la pertinence de leur suppression. Si ces variables sont redondantes avec d'autres, on choisit de les supprimer. Si elles peuvent avoir une importance d'un point de vue métier et que leur corrélation avec les autres variables semble être non justifiée intuitivement, on choisit de les conserver.

Par exemple, la variable VET605213 ("Veterans, 2009-2013") est liée (entre autres) à

- HSG010214 ("Housing units, 2014")
- SBO001207 ("Total number of firms, 2007").

Ces corrélations ne semblent pas directes intuitivement. De plus, aucune autre variable n'a un rapport avec les Vétérans. Ainsi, il convient de conserver cette variable malgré son VIF important.

Finalement, voici la liste des variables effectivement supprimées:
['RTN130207','BZA110213','BZA010213','RHI625214','Hispanic', 'INC110213','NonEnglish', 'RHI525214','NES010213','HSG010214','HSD410213','SBO001207','RHI425214','SBO115207','White']

2) Classification des comtés

a. **Analyse en composantes principales et interprétation :**

L'objectif de l'Analyse en Composantes Principales (ACP) est de générer un espace de dimension réduite en minimisant la perte de dispersion du nuage. Il s'agit donc d'obtenir le résumé le plus pertinent possible des données initiales.

Les composantes principales peuvent être considérées comme de nouvelles variables, combinaisons linéaires des variables initiales, non corrélées entre elles.

Nous utilisons SPAD pour générer et interpréter les composantes principales et voir quelles sont les variables qui contribuent le plus à celles-ci.

D'après les résultats de l'ACP, nous avons l'histogramme des valeurs propres suivants :

HISTOGRAMME DES 32 PREMIERES VALEURS PROPRES

NUMERO	VALEUR PROPRE	POURCENTAGE	POURCENTAGE CUMULE	
1	7.9756	24.92	24.92	**********
2	4.4741	13.98	38.91	**********
3	2.7520	8.60	47.51	**********
4	2.2135	6.92	54.42	**********
5	1.9126	5.98	60.40	**********
6	1.6956	5.30	65.70	**********
7	1.3549	4.23	69.93	**********
8	1.2955	4.05	73.98	**********
9	1.0122	3.16	77.14	**********
10	0.9593	3.00	80.14	**********
11	0.8805	2.75	82.89	**********
12	0.7424	2.32	85.21	**********
13	0.6554	2.05	87.26	**********
14	0.5672	1.77	89.03	**********
15	0.5125	1.60	90.64	**********
16	0.4149	1.30	91.93	**********
17	0.3411	1.07	93.00	**********
18	0.3044	0.95	93.95	**********
19	0.2779	0.87	94.82	**********
20	0.2367	0.74	95.56	**********
21	0.2162	0.68	96.23	**********
22	0.1945	0.61	96.84	**********
23	0.1772	0.55	97.39	**********
24	0.1590	0.50	97.89	**********
25	0.1456	0.46	98.35	**********
26	0.1260	0.39	98.74	**********
27	0.0996	0.31	99.05	**********
28	0.0898	0.28	99.33	**********
29	0.0753	0.24	99.57	**********
30	0.0609	0.19	99.76	**********
31	0.0562	0.18	99.93	**********
32	0.0214	0.07	100.00	**********

La "valeur propre" représente, pour chaque facteur, le montant de l'inertie du nuage sur ce facteur par rapport à la somme de toutes les valeurs propres qui représentent 100% de cette inertie. On prendra en compte un nombre limité de facteurs, en tenant compte les différences de pourcentages de variance expliquées par les facteurs.

Pout la sélection des axes factoriels à analyser nous avons choisi d'utiliser le **critère de Kaiser** qui consiste à retenir les axes pour lesquels la valeur propre est supérieure à la valeur propre moyenne.

Nous analyserons donc nos résultats sous l'angle des axes factoriels dont la valeur propre est supérieure à 1. On peut également utiliser le critère du coude : on retient les axes avant le décrochage. Les résultats de ces deux critères convergent : nous obtenons 9 facteurs.

L'étape suivante est l'interprétation des axes. Par souci de simplification, on va retenir seulement les 5 premiers axes factoriels (54,42% de l'inertie).

Interprétation des résultats

La colonne qui suit celle des coordonnées (CA ou contributions absolues de chaque élément) exprime la part prise par chaque élément d'une colonne du tableau dans la variance ou "inertie" expliquée par le facteur étudié). Ces contributions absolues permettent de savoir quels éléments sont responsables de la construction du facteur. La colonne CR (ou contribution relative ou corrélation) donne les valeurs prises par un facteur dans l'explication de la dispersion d'un élément. C'est la corrélation entre l'élément et l'axe factoriel considéré (elle varie donc entre 0 et 1 ou 0 et 1000, si les données sont en millièmes). La somme des CR d'un élément sur tous les facteurs est égale à 1 ou 1000.

VARIABLES ACTIVES

VARIABLES	COORDONNEES					CORRELATIONS VARIABLE-FACTEUR					ANCIENS AXES UNITAIRES				
IDEN - LIBELLE COURT	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
popu - population_change_ce	-0.57	0.15	-0.36	-0.20	-0.07	-0.57	0.15	-0.36	-0.20	-0.07	-0.20	0.07	-0.22	-0.13	-0.05
AGE1 - AGE135214_centrée-ré	-0.28	-0.52	-0.44	-0.40	0.09	-0.28	-0.52	-0.44	-0.40	0.09	-0.10	-0.25	-0.27	-0.27	0.06
AGE2 - AGE295214_centrée-ré	-0.18	-0.44	-0.39	-0.67	-0.02	-0.18	-0.44	-0.39	-0.67	-0.02	-0.06	-0.21	-0.23	-0.45	-0.02
age6 - age65plus_centrée-ré	0.47	0.41	0.45	0.22	0.11	0.47	0.41	0.45	0.22	0.11	0.17	0.19	0.27	0.15	0.08
SEX2 - SEX255214_centrée-ré	-0.15	0.04	-0.07	-0.09	-0.36	-0.15	0.04	-0.07	-0.09	-0.36	-0.05	0.02	-0.04	-0.06	-0.26
Blac - Black_centrée-réduit	-0.13	-0.63	0.06	0.42	-0.46	-0.13	-0.63	0.06	0.42	-0.46	-0.05	-0.30	0.03	0.28	-0.33
RHI3 - RHI325214_centrée-ré	-0.03	-0.09	-0.01	-0.08	0.53	-0.03	-0.09	-0.01	-0.08	0.53	-0.01	-0.04	-0.01	-0.05	0.38
RHI8 - RHI825214_centrée-ré	0.53	0.65	0.01	-0.24	0.06	0.53	0.65	0.01	-0.24	0.06	0.19	0.31	0.00	-0.16	0.04
POP7 - POP715213_centrée-ré	0.38	0.12	0.40	-0.41	-0.32	0.38	0.12	0.40	-0.41	-0.32	0.14	0.05	0.24	-0.27	-0.23
POP6 - POP645213_centrée-ré	-0.76	-0.10	-0.04	-0.09	0.22	-0.76	-0.10	-0.04	-0.09	0.22	-0.27	-0.05	-0.03	-0.06	0.16
Edu_ - Edu_highschool_centr	-0.16	0.81	-0.23	0.03	-0.06	-0.16	0.81	-0.23	0.03	-0.06	-0.06	0.38	-0.14	0.02	-0.04
Edu_ - Edu_batchelors_centr	-0.64	0.53	-0.31	0.15	-0.12	-0.64	0.53	-0.31	0.15	-0.12	-0.23	0.25	-0.19	0.10	-0.09
VET6 - VET605213_centrée-ré	-0.80	0.05	0.46	-0.04	0.03	-0.80	0.05	0.46	-0.04	0.03	-0.28	0.02	0.28	-0.03	0.02
LFE3 - LFE305213_centrée-ré	-0.06	-0.15	0.31	-0.33	-0.46	-0.06	-0.15	0.31	-0.33	-0.46	-0.02	-0.07	0.19	-0.22	-0.33
HSG4 - HSG445213_centrée-ré	0.58	0.40	0.28	-0.46	-0.19	0.58	0.40	0.28	-0.46	-0.19	0.20	0.19	0.17	-0.31	-0.13
HSG0 - HSG096213_centrée-ré	-0.76	0.13	-0.24	0.36	0.01	-0.76	0.13	-0.24	0.36	0.01	-0.27	0.06	-0.15	0.24	0.01
HSG4 - HSG495213_centrée-ré	-0.66	0.40	-0.15	0.00	-0.08	-0.66	0.40	-0.15	0.00	-0.08	-0.23	0.19	-0.09	0.00	-0.06
HSD3 - HSD310213_centrée-ré	-0.39	-0.55	-0.18	-0.52	-0.05	-0.39	-0.55	-0.18	-0.52	-0.05	-0.14	-0.26	-0.11	-0.35	-0.04
Inco - Income_centrée-rédui	-0.49	0.71	-0.19	-0.09	-0.22	-0.49	0.71	-0.19	-0.09	-0.22	-0.18	0.34	-0.12	-0.06	-0.16
Pove - Poverty_centrée-rédu	0.14	-0.75	0.17	0.37	0.09	0.14	-0.75	0.17	0.37	0.09	0.05	-0.36	0.10	0.25	0.06
BZA1 - BZA115213_centrée-ré	-0.11	0.01	-0.03	-0.11	-0.14	-0.11	0.01	-0.03	-0.11	-0.14	-0.04	0.01	-0.02	-0.08	-0.10
SBO3 - SBO315207_centrée-ré	-0.26	-0.40	-0.03	0.32	-0.49	-0.26	-0.40	-0.03	0.32	-0.49	-0.09	-0.19	-0.02	0.21	-0.35
SBO2 - SBO215207_centrée-ré	-0.65	0.01	0.06	0.00	-0.08	-0.65	0.01	0.06	0.00	-0.08	-0.23	0.00	0.03	0.00	-0.06
SBO5 - SBO515207_centrée-ré	-0.17	0.02	-0.05	0.01	-0.03	-0.17	0.02	-0.05	0.01	-0.03	-0.06	0.01	-0.03	0.01	-0.02
SBO4 - SBO415207_centrée-ré	-0.55	-0.21	0.07	-0.14	0.28	-0.55	-0.21	0.07	-0.14	0.28	-0.20	-0.10	0.04	-0.10	0.21
SBO0 - SBO015207_centrée-ré	-0.43	-0.06	-0.22	0.10	-0.21	-0.43	-0.06	-0.22	0.10	-0.21	-0.15	-0.03	-0.14	0.07	-0.15
MAN4 - MAN450207_centrée-ré	-0.68	0.02	0.51	-0.10	-0.01	-0.68	0.02	0.51	-0.10	-0.01	-0.24	0.01	0.31	-0.07	-0.01
WIN2 - WIN220207_centrée-ré	-0.70	0.03	0.55	-0.09	0.01	-0.70	0.03	0.55	-0.09	0.01	-0.25	0.01	0.33	-0.06	0.01
RTN1 - RTN131207_centrée-ré	-0.45	0.28	-0.34	0.21	0.05	-0.45	0.28	-0.34	0.21	0.05	-0.16	0.13	-0.20	0.14	0.04
AFN1 - AFN120207_centrée-ré	-0.79	0.04	0.52	-0.03	0.03	-0.79	0.04	0.52	-0.03	0.03	-0.28	0.02	0.31	-0.02	0.02
BPS0 - BPS030214_centrée-ré	-0.78	0.07	0.41	-0.07	0.03	-0.78	0.07	0.41	-0.07	0.03	-0.28	0.03	0.24	-0.05	0.02
LND1 - LND110210_centrée-ré	-0.23	-0.04	0.10	-0.03	0.62	-0.23	-0.04	0.10	-0.03	0.62	-0.08	-0.02	0.06	-0.02	0.45

Analyse des résultats sur le premier axe factoriel

On voit que les variables :

- HSG445213 (Homeownership rate 2009-2013 - 0.58)
- RHI825214 (White alone, not Hispanic or Latino, percent, 2014 - 0.53)

ont une valeur positive proche de 1. Ces variables contribuent positivement à la construction de l'axe 1.

A l'inverse les variables :

- VET605213(-0.80) : Veterans, 2009-2013
- HSG096213(-0.76) Housing units in multi-unit structures, percent, 2009-2013
- POP645213 (-0.76) : Foreign born persons, percent, 2009-2013 ont une valeur proche de -1.

Ces variables contribuent négativement à la construction de l'axe1.

Ce premier axe oppose donc :

- les comtés à majorité blanche avec une proportion de propriétaires élevée et un faible nombre d'étrangers
- avec les comtés à faible proportion de Blancs, une forte proportion d'étrangers et une faible proportion de propriétaires

Analyse des résultats sur le deuxième axe factoriel

Les variables Edu_highschool (0.80) et Edu_bachelors (0.53) contribuent positivement à la construction de l'axe 2, contrairement à la variable Poverty(-0.75) qui y contribue négativement.

On peut donc dire que le deuxième axe factoriel oppose :

- les comtés avec un fort niveau d'éducation et un fort niveau de richesse
- des comtés avec un faible niveau d'éducation et de richesse

Analyse des résultats sur le troisième axe factoriel

Les variables

- WTN220207 (0.55) : Merchant wholesaler sales, 2007 ($1,000)
- POP715213 : "Living in same house 1 year & over, percent, 2009-2013" (0.40)
- age65plus (0.45)

contribuent positivement à la construction de l'axe 3.

Par contre les variables :

- AGE135214 (-0.44) : Persons under 5 years, percent, 2014
- AGE295214 (-0.39) : Persons under 18 years, percent, 2014

y contribuent négativement.

On conclue donc que l'axe 3 oppose les comtés :

- Avec une faible proportion de jeunes et des personnes assez sédentaires
- Et ceux avec une forte proportion de jeunes et une population plus mobile

Analyse des résultats sur le quatrième axe factoriel

Les variables Black (0.42) et Poverty (0.37) contribuent à la construction de l'axe 4.

A l'inverse, la variable AGE295214 (-0.67) Persons under 18 years, percent, 2014 y contribuent négativement.

L'axe 4 oppose probablement les comtés relativement pauvres avec une population à majorité noire et âgée des comtés plus riches et plus jeunes avec une moins forte représentation des Noirs.

Analyse des résultats sur le cinquième axe factoriel

Les variables :

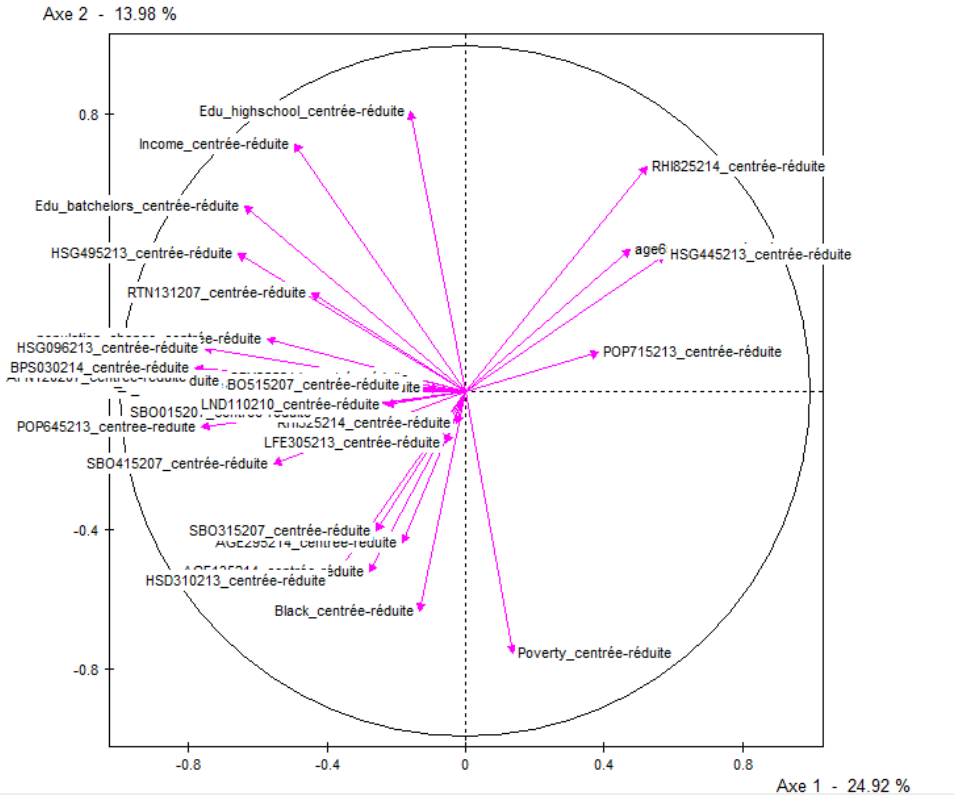
- RHI325214 (0.53) : American Indian and Alaska Native alone, percent, 2014
- LND110210 (0.62) : Land area in square miles, 2010

contribuent positivement à la construction de l'axe 5 contrairement à la variable SBO315207 (-0.49) : Black-owned firms, percent, 2007

Donc l'axe 5 oppose les comtés sur leur superficie, la proportion d'Indiens dans la population ainsi que le taux d'entreprises détenues par des Noirs d'Amérique.

Représentation graphique :

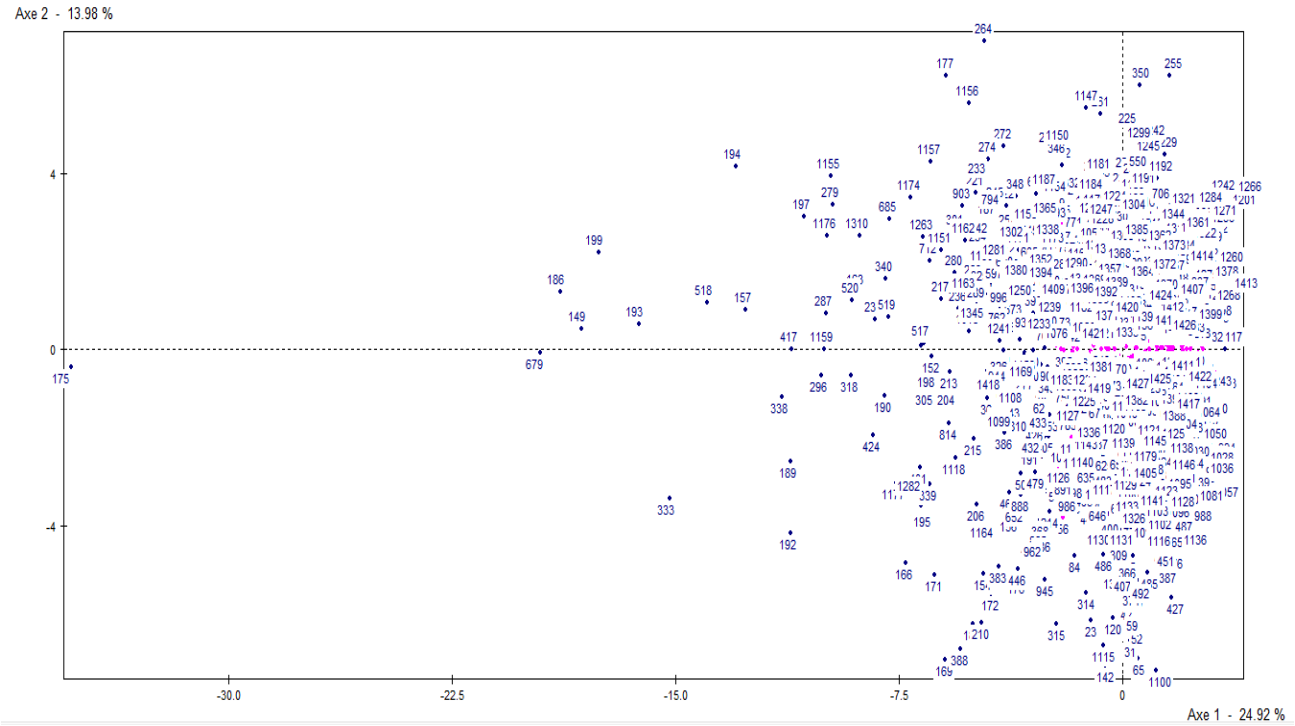
Nous avons projeté les résultats de cette ACP sur le premier plan factoriel (deux premiers axes) afin d'illustrer visuellement les conclusions tirées ci-dessus. Ces deux axes expliquent 38.9% d'information (inertie).



Graphiquement, on voit bien les variables positives sont bien représentées et proches du cercle de corrélation, idem pour les variables qui se retrouvent du côté négatif citées plus haut.

<u>Représentation graphique des individus</u>

On a projeté les individus statistiques (ID : comtés) sur le premier plan factoriel.
Les comtés proches des extrémités de chaque axe sont celles qui sont le mieux représentées sur ceux-ci, et donc celles sur lesquelles on peut effectuer une interprétation des résultats.



La distribution des comtés sur l'axe 1-2 est très concentrée. Les comtés sont représentés par des ID uniques allant de 1 à 3112.

Interprétation des individus avec pour référentiel l'axe 1

Par exemple, les individus numéros 1266, 1413 , 1260, 1268,1378 sont du côté positif de l'axe 1, ils représentent les comtés suivants rangés par ordre : St-Lucie, Ogemaw, Henry, Osceola et Bates.

Ces comtés appartiennent aux états suivants : Michigan, Florida, Michigan, Michigan et le Montana, et sont caractérisés majoritairement par des populations blanches, propriétaires de leurs maisons et plutôt âgées. Le Michigan est très représentatif dans cette partie de l'ACP.

Les individus du côté négatif sont représentés par les ID suivants : 988, 488, 1136 sont les comtés Clark, Talbot et Tangipahoa Parish sont caractérisés par une forte proportion de vétérans étrangers. Ils appartiennent aux états suivants : Kentucky, Georgie et la Louisiane. Ces états ont des populations plutôt étrangères, composées de vétérans ayant des logements unitaires.

Interprétation des individus avec pour référentiel l'axe 2

Les individus du cotés positif de l'axe 2 (264, 177, 1156, 1147, 350 et 255) correspondent aux comtés suivants : Phillips, Madera, Middlesex, Winn Parish, St, Lucie et Mesa.

Ces comtés sont dans les Etats : Colorado, Californie, Massachusetts, Louisiane, Floride et Colorado.

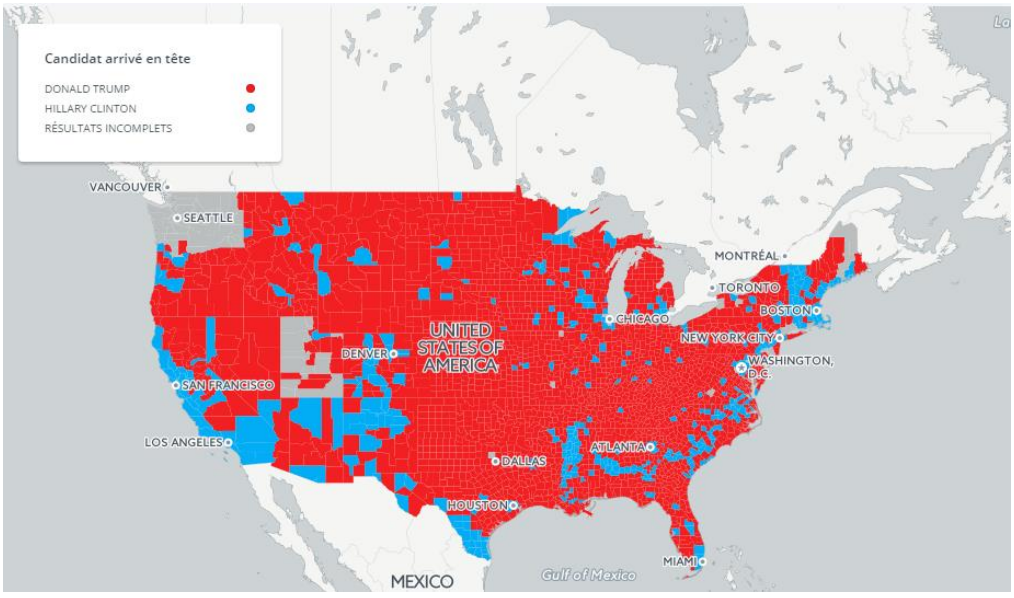
Ces comtés ont des caractéristiques très fortes en termes de niveau d'éducation de la population. Les gens qui y vivent sont à moitié propriétaires et ont un âge supérieur à 65 ans, ce sont donc des états riches avec un revenu supérieur à la moyenne.

Les comtés qui se trouvent proches du côté négatif de l'axe 2 (110, 169, 142, 388, 1115, 315 et 1210) sont Lonoke, Humbold, Yell, Clay, Livingston Parish, Hardee et Bay. Ces comtés appartiennent aux états suivants : Arkansas, Californie, Arkansas, Georgie, Louisiane, Floride et le Michigan. Ces comtés sont très pauvres avec une forte population noire. Ils se caractérisent par des personnes plutôt jeunes (- de 18 et 5 ans) nées à l'étranger.

Cette ACP permet de nous éclairer sur les différentes caractéristiques de chaque comté selon les indicateurs économiques disponibles dans notre base.

Elle nous donne des informations très précises sur les variables corrélées, les variables contribuant à la construction des axes fournissant la meilleure inertie mais surtout une vision globale qui confirmera notre typologie et classifications de nos variables par rapport aux individus. Les variables corrélées correspondent exactement aux variables relevées plus haut dans la partie des corrélations. Nous remarquons que toutes les corrélations linéaires sont positives et négatives sur les 5 axes facteurs en considérant toutes les variables (ce qui signifie que toutes les variables positives et négatives varient, en moyenne, dans un même sens), certaines étant moyennes (0.68 et 0.53 et -0.55 et -0.65), d'autres plutôt faibles.

Ci-dessous un graphique représentant les votes des deux candidats Clinton et Trump selon chaque comté. Source : http://www.francetvinfo.fr/monde/usa/presidentielle/carte-presidentielle-americaine-decouvrez-tous-les-resultats-comte-par-comte_1914633.html



b. **Mise en place d'une classification ascendante hiérarchique**

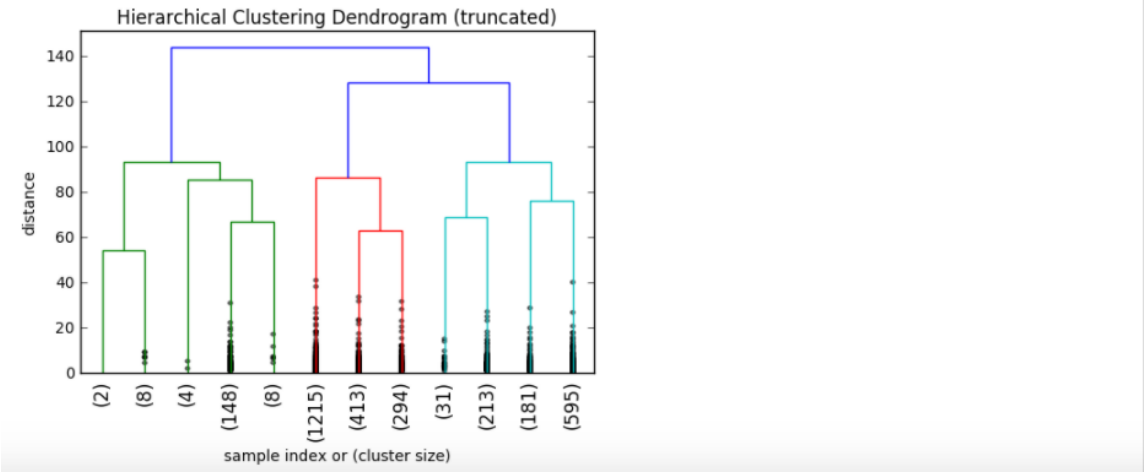
L'ACP nous a donc permis de générer un nombre plus réduit de variables non corrélées. On s'appuie donc sur ces nouveaux axes factoriels pour réaliser la classification des comtés.

A cet égard, la classification ascendante hiérarchique semble être une méthode pertinente. On choisit également le critère de Ward qui regroupe les différents comtés en minimisant la perte d'inertie interclasse (ou en maximisant l'inertie intra-classe).

La CAH nous donne le dendrogramme suivant :

```
#Construction du dendrogramme

plt.title('Hierarchical Clustering Dendrogram (truncated)')
plt.xlabel('sample index or (cluster size)')
plt.ylabel('distance')
dendrogram(
    Z,
    truncate_mode='lastp',  # show only the last p merged clusters
    p=12,  # show only the last p merged clusters
    leaf_rotation=90.,
    leaf_font_size=12.,
    show_contracted=True,  # to get a distribution impression in truncated branches
)
plt.show()
```



L'enjeu est ici de couper le dendrogramme à l'endroit qui nous semble le plus pertinent. On remarque qu'une découpe au seuil 100 est cohérente :

- Tout d'abord, cela nous permet de ne pas démultiplier le nombre de classes et de rester à un niveau macro. Si cela est nécessaire, nous pourrons rentrer plus dans le détail par la suite et générer plus de classes pour gagner en précision en termes d'interprétation des résultats.

- De plus, la perte d'inertie intraclasse entre 3 classes et 2 classes nous pousse également à couper le dendrogramme à ce seuil.

Il n'en reste pas moins que la pertinence de la découpe dépend de l'analyse des classes. En effet, si les classes semblent homogènes en termes de caractéristiques socio-économiques, la CAH pourra être validée.

c. Interprétation des classes

Tout d'abord, il apparaît important de calculer le nombre de comtés par classes. La répartition est la suivante :

- Classe 1: 170
- Classe 2: 1922
- Classe 3: 1020

Les comtés de la 1ère classe et il semble pertinent de les isoler et de les traiter séparément, bien que le volume soit très faible par rapport aux deux autres classes. Dans un 2e temps, il pourrait être intéressant de scinder la classe 2 en plusieurs sous-classes.

La répartition des votes nous donne également un indice sur l'homogénéité des classes :

- 71% des comtés de la classe 1 ont assisté à la victoire de Clinton en 2016
- 9% dans la classe 2
- 18% dans la classe 3

Cela est plutôt homogène. L'analyse des classes nous permettra d'ailleurs déjà de comprendre les motifs des différents votes.

Afin de déterminer les principales différences entre chaque classe, nous nous sommes focalisés sur les moyennes et les écarts-types de chacune des classes pour chaque variable.

Grâce à des tests de Student (two-sample t-tests), nous avons retenu les variables pour lesquelles l'écart entre les moyennes des classes est très significatif (et les écart-types globalement identiques).

In [45]:
```
#Tests statistiques de Student pour tester si les moyennes sont bien différentes
student_tests = []
for i in variables_discriminantes:
    tval, pval = stats.ttest_ind(X_center.ix[X_center['Class'] == 1,i],X_center.ix[X_center['Class'] == 2,i], equal_var
    if i in dictionnaire_variables:
        student_tests.append((i,dictionnaire_variables[i],int(tval), pval))

for elt in student_tests:
    print(elt)
```

```
('MAN450207', '"Manufacturers shipments, 2007 ($1,000)"', 8, 4.160238023855220 8e-14)
('WTN220207', '"Merchant wholesaler sales, 2007 ($1,000)"', 7, 1.4372655533189593e-12)
('BPS030214', '"Building permits, 2014"', 9, 2.5652928261042814e-17)
('HSG495213', '"Median value of owner-occupied housing units, 2009-2013"', 10, 3.3256840876215538e-21)
('VET605213', '"Veterans, 2009-2013"', 13, 1.0042347474707208e-29)
('POP645213', '"Foreign born persons, percent, 2009-2013"', 17, 4.2861796576785038e-39)
('SBO215207', '"Asian-owned firms, percent, 2007"', 11, 1.713940608185916e-23)
('AFN120207', '"Accommodation and food services sales, 2007 ($1,000)"', 8, 5.631216452499922 2e-15)
('HSG096213', '"Housing units in multi-unit structures, percent, 2009-2013"', 16, 1.1804817399773908e-37)
```

Ce bout de script vérifie que la différence de moyennes pour chaque variable discriminante est bien significative. On voit ici que la p-value est très inférieur à 5% dans chacun des cas. Si cela n'est pas le cas, la variable est ignorée pour caractériser les classes.

Les variables discriminantes retenues nous ont permis de définir les classes relativement les unes par rapport aux autres. En guise d'exemples, les 3 tableaux ci-dessous montrent quelques-unes de ces variables discriminantes.

```
#Visualisation des moyennes prises par les différentes variables différenciantes
X_center[X_center['Class'] == 1].describe()[variables_discriminantes].filter(['mean','std'],axis=0)
```

	VET605213	MAN450207	Edu_batchelors	Density	Edu_highschool	Black	Income	HSG096213	SBO215207	SBO415207	WTN220207
mean	2.791346	2.08452	1.565590	1.368873	0.442071	0.381675	1.376912	1.990726	2.418068	0.914601	2.039688
std	2.755009	3.47277	1.060013	3.930800	0.570935	0.680738	1.293565	1.569496	2.857726	1.272704	3.674801

```
X_center[X_center['Class'] == 2].describe()[variables_discriminantes].filter(['mean','std'],axis=0)
```

	VET605213	MAN450207	Edu_batchelors	Density	Edu_highschool	Black	Income	HSG096213	SBO215207	SBO415207	WTN2202(
mean	-0.141867	-0.114697	0.198384	-0.076908	0.486571	-0.402564	0.286017	-0.018925	-0.143432	-0.161066	-0.117131
std	0.371486	0.290680	0.941322	0.233188	0.632077	0.365373	0.862981	0.855530	0.438684	0.376635	0.139043

```
X_center[X_center['Class'] == 3].describe()[variables_discriminantes].filter(['mean','std'],axis=0)
```

	VET605213	MAN450207	Edu_batchelors	Density	Edu_highschool	Black	Income	HSG096213	SBO215207	SBO415207	WTN2202(
mean	-0.197902	-0.131295	-0.634748	-0.083227	-0.990531	0.694944	-0.768431	-0.296127	-0.132741	0.151065	-0.119236
std	0.396216	0.355755	0.556561	0.239213	0.885717	1.382418	0.556648	0.717658	0.551248	1.518423	0.206689

On voit par exemple que la variable VET605213 (Veterans, 2009-2013) est discriminante, dans le sens où les valeurs moyennes de chaque classe pour cette variable sont très différentes. On peut conclure que la classe 1 regroupe des comtés dans lesquels, en moyenne, la part des vétérans est bien plus fort que dans les comtés des classes 2 et 3.

En fonctionnant de manière similaire pour toutes les variables discriminantes, on peut dresser un descriptif rapide de chaque classe (comparativement aux autres classes)

La classe 1 regroupe plutôt des comtés avec :

- un niveau d'éducation élevé
- un revenu moyen par habitant élevé
- une densité de population forte (environnement urbain)
- une forte proportion d'étrangers (Asiatiques, Hispaniques et Noirs)
- une forte activité économique
- un marché de l'immobilier actif et cher

On a donc affaire ici à des comtés riches, urbains, éduqués et cosmopolites. On remarque que c'est le genre de contexte socio-économique dans lequel les positions de la candidate démocrate ont de l'écho. Cela est appuyé par le fait que plus de 70% de ces 170 comtés ont voté à majorité Clinton.

La classe 2, de son côté, comprend des comtés avec :

- un niveau d'éducation et de revenu par habitant dans la moyenne
- une proportion d'étrangers en-dessous de la moyenne
- une proportion de blancs très importante
- un taux de propriétaire au-dessus de la moyenne

La classe 2 comprend par conséquent des comtés à caractère plus rural avec une faible proportion d'étrangers et une surreprésentation des Blancs. C'est en quelque sorte une des cibles sur lesquelles se concentrer Donald Trump.

La classe 3, enfin, est constitué de comtés avec :

- un faible niveau d'éducation et de revenu
- une sévère pauvreté
- une densité de population faible
- un marché de l'immobilier très bas
- une proportion importante d'étrangers (en particulier de Noirs d'Amérique)

C'est également une classe dans laquelle le discours de Donald Trump a fédéré (moins de 20% de victoires pour Clinton dans cette classe). En effet, Trump est arrivé à séduire ces citoyens américains faiblement éduqués et oubliés de la mondialisation. Bien entendu les citoyens blancs, mais également les étrangers, ce qui peut paraître surprenant vu ses positions extrêmes sur certains sujets tels que l'immigration.

Cette typologie nous permet de distinguer des sous-populations homogènes dans les comtés américains et de commencer à analyser les résultats selon les caractéristiques de chaque classe.

L'enjeu est maintenant d'aller plus loin dans l'explication des votes en entraînant des régressions multiples pénalisées sur chacune des classes.

3) Régressions par classes de comtés

a. Procédure ElasticNet

Etant donné le nombre important de variables potentiellement explicatives en jeu (35), nous avons décidé de rajouter une pénalisation aux régressions appliquées. La pénalisation choisie est ElasticNet, c'est une méthode de régression régularisée qui combine linéairement les pénalités L1 et L2 des méthodes de lasso et ridge, il résout les limites des deux méthodes, tout en incluant chacune comme cas particulier. Donc, si ridge ou la solution de lasso est, en effet, le meilleur, alors toute bonne sélection de modèle de routine permettra d'identifier cela dans le cadre du processus de modélisation.

L'intuition mathématique derrière ce modèle est la suivante :

$$\textbf{Residual Mean Square Error} + \alpha \cdot \textbf{Ridge Penalty} + (1-\alpha) \cdot \textbf{LASSO}$$

pour $\alpha \in [0,1]$.

Interprétation de la régression de Clinton sur la classe 1 :

Dans la classe 1, le bon modèle retient 18 variables explicatives.

Ci-dessous les résultats de la régression ElasticNet sur la variable cible Clinton:

	Coefficients	Variables	Descriptif
0	4,66935958	per_point_diff_2012	
15	-1,41762965	HSG445213	"Homeownership rate, 2009-2013"
17	1,19040886	HSG495213	"Median value of owner-occupied housing units, 2009-2013"
12	1,07467126	Edu_batchelors	
10	0,72411962	POP645213	"Foreign born persons, percent, 2009-2013"
22	0,66668466	SBO315207	"Black-owned firms, percent, 2007"
16	0,54189862	HSG096213	"Housing units in multi-unit structures, percent, 2009-2013"
8	-0,48276205	RHI825214	"White alone, not Hispanic or Latino, percent, 2014"
3	-0,44643428	AGE295214	"Persons under 18 years, percent, 2014"
6	0,37105752	Black	
18	-0,25696393	HSD310213	"Persons per household, 2009-2013"
27	0,23689953	MAN450207	"Manufacturers shipments, 2007 ($1,000)"
9	-0,18174444	POP715213	"Living in same house 1 year & over, percent, 2009-2013"
31	-0,09811194	BPS030214	"Building permits, 2014"
20	0,08720505	Poverty	
19	0,06822065	Income	
28	-0,04922661	WTN220207	"Merchant wholesaler sales, 2007 ($1,000)"
13	-0,00803104	VET605213	"Veterans, 2009-2013"

La classe 1 regroupe les comtés suivants :

0	0
15	Cleburne
17	Colbert
12	Choctaw
10	Cherokee
22	Cullman
16	Coffee
8	Calhoun
3	Barbour
6	Bullock
18	Conechuh
27	Escambia
9	Chambers
31	Geneva
20	Covington
19	Coosa
28	Etowah
13	Clarke

Les variables qui font partie de cette classe sont expliquées par rapport au coefficients disponibles dans le tableau.

Elastic Net a donc choisi le bon modèle relatif à la première classe 1. Lors d'une régression multiple, les coefficients (paramètre estimés de notre modèle) influent sur notre variable cible si nous l'augmentons ou diminuons d'une unité.

Par rapport aux votes des élections de 2012, nous constatons quatre points de différence avec ceux des élections de 2016. Il y a donc eu plus de votes en Alabama en 2016 qu'en 2012.

Sachant que les plus grands coefficients (=> à 1) de la classe 1 sont HSG495213 : « Valeur médiane des logements occupés par leur propriétaire, 2009-2013» et Edu_bachelors, on voit que ces deux variables expliquent bien les votes de la candidate Clinton aux élections, on retient que finalement les gens instruits et diplômés préfèrent Clinton à Trump. En plus de la population née à l'étranger, des chefs d'entreprises noirs et des propriétaires de maisons.

Si l'on augmente nos variables HSG495213 et Edu_bachelors d'une unité, nous nous attendons à ce que notre variable cible Clinton augmentera également de 1.19 et 1.07. Ces coefficients changent notre Y si l'on change nos variables X d'une unité.

a. Procédure 'Stepwise'

Un trop grand nombre de variables explicatives relativement au nombre d'observations peut entrainer une diminution de la précision de l'estimation des coefficients des variables explicatives et une diminution du nombre de degré de liberté de la résiduelle et donc potentiellement une augmentation de la variance résiduelle. Il est ainsi nécessaire de retirer les variables qui ne contribuent pas ou très peu à expliquer la variable cible et celles qui apportent une information redondante.

On cherchera donc à privilégier le modèle qui explique le mieux la variable d'intérêt aven un minimum de variables explicatives.

Il existe un grand nombre pour comparer des modèles entre eux. Nous utilisons dans cette étude le critère AIC (Akaike Information Criterium) défini comme :

$$AIC = 2k-2ln(L)$$

Avec k, le nombre de paramètres du modèle et L la vraisemblance des données observées sous le modèle, c'est à dire pour les variables continues elle est égale à la valeur de la densité de probabilité des données observées dans le modèle. Nous sélectionnerons donc le modèle avec l'AIC le plus petit (minimisation).

Les méthodes les plus souvent utilisées en sélection des variables sont des méthodes heuristiques, des méthodes numériques qui fournissent une solution rapide mais pas nécessairement optimale, comme la procédure ascendante (Forward) ou descendante (Backword). Stepwise intègre le critère AIC et représente la procédure ou on commence avec le modèle le plus complexe, le modèle complet, celui incluant toutes les variables. On calcule son AIC, et à l'étape suivante, on définit des 'Nouveaux' modèles, en enlevant une de ces variables et en ajoutant. On compare l'AIC de chacun de ces nouveaux modèles avec l'AIC précédant, si aucun AIC n'est mieux que le précédant, on stoppe la procédure, sinon on continue avec le nouveau modèle et on recommence ce processus jusqu'à obtenir le meilleur modèle.

b. Analyse des résultats

Nous avons utilisé les résultats de classification et essayé de trouver les meilleurs modèles pour chaque classe. Nous partons d'une hypothèse telle que dans chaque classe de comptés, il y ait des caractéristiques différentes qui contribuent à l'explication des résultats des votes pour Clinton et Trump.

Suite à la procédure Stepwise, nous retenons les modèles suivants :

Class 1

Trump*= per_point_diff_2012 +HSG445213+ Edu_batchelors + POP645213 + HSG096213 + HSD310213 + MAN450207 + POP715213 + Poverty + Income + WTN220207*

Clinton*= per_point_diff_2012 + HSG445213 + Edu_batchelors + POP645213 + SBO315207+HSG096213 + HSD310213 + Poverty + Income*

On peut remarquer qu'on observe presque les mêmes variables pour les deux candidats, c'est-à-dire dans ce compté dominent des facteurs comme : le taux d'accession à la propriété, le revenu, le pourcentage des gens avec un niveau d'éducation 'batchelor' etc. Si on regarde les caractéristiques de cette classe on constate que cela correspond bien aux comptés riches, urbains, éduqués et cosmopolites.

Cependant, pour Clinton on a SBO315207 qui représente le pourcentage des entreprises qui appartiennent aux personnes 'noirs' et dans le modèle de Trump on a des variables 'uniques' comme les ventes en gros, le pourcentage des gens qui habitent dans la même maison plus d'un an et le nombre des expéditions manufacturières.

Réalisons le test global du modèle dans lequel les hypothèses testées sont les suivantes :

H0 : $Yi= a + Ei$, $Ei \sim N(0,\sigma^2)$

H1: $Yi= a + \Sigma\ bk * xik + Ei$, $Ei \sim N(0,\sigma^2)$

Les p-value pour Clinton et Trump sont inférieure à 0,05. Donc au risque de 5% on rejette H0, ce qui signifie qu'il existe au moins une influence d'une des caractéristiques des comptés sur les résultats des votes.

En examinant les estimations des paramètres du modèle pour Trump, nous nous apercevons que toutes les variables sauf WTN220207 et Income ont une influence statistiquement significative au risque de 5%. Notamment, on constate qu'il existe un lien positif entre la différence per points en 2012, le pourcentage des gens avec 'Bachelor degree', les niveaux de pauvreté et le pourcentage des votes pour Trump. Ce sont des effets très intéressants, car la différence par point des votes est grande dans les comtés ou Obama a gagné en 2012,c'est-à-dire qu'il existe une sorte de mécontentement avec la politique d'Obama d'où le vote dirigé vers le parti républicain, car finalement 71% des comtés de la classe 1 ont assisté à la victoire de Clinton en 2016. Il y a quand même une minorité de la population qui a voté pour Trump malgré la majorité ayant voté pour Clinton. Dans cette classe on a le coefficient pour cette variable le plus forte en comparaison avec des autres classes.

Tableau 1 Trump Classe 1

Estimations par l'analyse du maximum de vraisemblance						
Paramètre	DDL	Valeur estimée des paramètres	Erreur type	Khi-2	Pr > Khi-2	Rapport de risque
per_point_diff_2012	1	6.42404	0.46331	192.2532	<.0001	616.488
HSG445213	1	-0.98828	0.25429	15.1038	0.0001	0.372
Edu_batchelors	1	1.44942	0.21984	43.4687	<.0001	4.261
POP645213	1	0.23222	0.11619	3.9947	0.0456	1.261

Estimations par l'analyse du maximum de vraisemblance						
Paramètre	DDL	Valeur estimée des paramètres	Erreur type	Khi-2	Pr > Khi-2	Rapport de risque
HSG096213	1	-0.81259	0.16575	24.0354	<.0001	0.444
HSD310213	1	0.38295	0.16032	5.7058	0.0169	1.467
MAN450207	1	0.11987	0.05416	4.8987	0.0269	1.127
POP715213	1	-0.69631	0.16448	17.9225	<.0001	0.498
Poverty	1	0.82992	0.27609	9.0360	0.0026	2.293
Income	1	0.36908	0.20851	3.1332	0.0767	1.446
WTN220207	1	-0.08019	0.05479	2.1421	0.1433	0.923

Pour Clinton, on a les variables « *per_point_diff_2012, HSG445213, Edu_batchelors , POP645213, SBO315207, HSG096213, HSD310213, Poverty* » qui sont significatif. En ce qui concerne les coefficients ici on observe une situation logiquement opposée. Ainsi la différence par points en 2012 et le pourcentage des gens avec 'Bachelor degree', le niveau de pauvreté ont un lien négatif avec le pourcentage des votes pour Clinton. De plus, le grand pourcentage de personnes nées à l'étranger et le nombre des gens par ménage ne jouent pas un rôle positif dans la distribution des votes pour Clinton. En fait, les caractéristiques qui reflètent négativement une image des comtés 'Elite' comme celle-ci, contribuent à la diminution des votes pour Clinton dans les comtés ou apriori elle était numéro 1.

Tableau 2 Clinton Classe 1

Estimations par l'analyse du maximum de vraisemblance						
Paramètre	**DDL**	**Valeur estimée des paramètres**	**Erreur type**	**Khi-2**	**Pr > Khi-2**	**Rapport de risque**
per_point_diff_2012	1	-6.91172	0.52830	171.1655	<.0001	0.001
HSG445213	1	0.73955	0.18081	16.7299	<.0001	2.095
Edu_batchelors	1	-1.34832	0.21868	38.0153	<.0001	0.260
POP645213	1	-0.35900	0.11396	9.9241	0.0016	0.698
SBO315207	1	-0.32571	0.12047	7.3095	0.0069	0.722
HSG096213	1	0.51613	0.11944	18.6742	<.0001	1.676
HSD310213	1	-0.47597	0.17438	7.4500	0.0063	0.621
Poverty	1	-1.16707	0.33417	12.1968	0.0005	0.311
Income	1	-0.35692	0.19599	3.3164	0.0686	0.700

Classe 2

En ce qui concerne la classe deux on a obtenu les modèles suivants :

Trump*=per_point_diff_2012 +HSG445213 + HSG495213 + Edu_batchelors + POP 645213 +HSG096213 +*
RHI825214 + AGE295214 + Black + BPS030214 + Poverty + Income

Clinton*=per_point_diff_2012+ HSG495213 +Edu_batchelors+ POP645213+ HSG 096213 RHI825214 + AGE295214+ Black+ HSD310213+POP715213+BPS030214+Povert y+Income*

Comme pour la classe 1 on constate que les variables sorties pas la procédure Stepwise sont Presque identiques pour les deux candidats. Notamment, que le taux d'accession à la propriété, le nombre de personnes par ménages et le pourcentage de gens qui habitent dans la même maison, appartiennent au modèle de Trump.

Comme on l'a déjà défini la classe 2 comprend des comtés plutôt ruraux avec une faible proportion d'étrangers et une surreprésentation des Blancs, on peut dire les 'comptés classiques blancs'. Dans ces deux modèles on voit bien que parmi les caractéristiques choisies par Stepwise on a la valeur médiane des unités de logement occupées par le propriétaire, le nombre d'autorisations de construire, le pourcentage des personnes noires, d'origine latino-américaine et blancs.

Réalisons le test global du modèle on remarque que les p-value pour Clinton et Trump sont inférieure à 0,05, dons on peut dire qu'au risque de 5% on rejette H0, et qu'il existe au moins une influence d'un des caractéristiques des comptés sur les résultats des votes dans cette classe.

En regardant sur les résultats de test de significativité de chaque paramètre, on constate que toutes les variables du modèle pour Trump ont une influence statistiquement significative au risque de 5% sur ses résultats. On voit le même effet positif avec les variables la différence per points en 2012, le pourcentage de gens avec 'Bachelor degree' comme dans la classe 1, par contre la variable 'poverty' dans cette classe est liées négativement avec le pourcentage des votes pour Trump, ainsi que le revenu, qui n'était pas significatif pour les 'comptés elite'. Ici les résultats de Trump sont positivement liés avec le pourcentage de population noires. C'est-à-dire dans ces comptés dominés par les blancs, la variété de races est une situation favorable pour la compagne de Trump . N'oublions pas que ces comptés sont une cible sur lesquelles Donald Trump a décidé de se concentrer. En regardant sur le lien négatif entre le niveau de pauvreté et le niveau de revenu avec les votes pour Trump, on peut constater qu'il cherche le support des américains dits d'une classe 'moyenne' qui ne sont pas riche mais qui ne sont pas non plus très pauvre, qui habitent dans des endroits avec une population variée. Comme ça il peut jouer avec sa politique contre les immigrés et cibler les blancs 'extrêmes'.

Tableau 3 Trump Classe 2

Estimations par l'analyse du maximum de vraisemblance						
Paramètre	DDL	Valeur estimée des paramètres	Erreur type	Khi-2	Pr > Khi-2	Rapport de risque
per_point_diff_2012	1	4.55043	0.19851	525.4490	<.0001	94.673
HSG445213	1	0.29053	0.08443	11.8417	0.0006	1.337
HSG495213	1	1.25257	0.13524	85.7820	<.0001	3.499
Edu_batchelors	1	0.75733	0.12297	37.9306	<.0001	2.133
POP645213	1	0.19944	0.06065	10.8132	0.0010	1.221
HSG096213	1	0.39306	0.12742	9.5161	0.0020	1.482
RHI825214	1	-0.66999	0.10581	40.0936	<.0001	0.512
AGE295214	1	0.22845	0.05396	17.9222	<.0001	1.257
Black	1	0.43496	0.05830	55.6561	<.0001	1.545
BPS030214	1	0.40692	0.16035	6.4404	0.0112	1.502
Poverty	1	-0.26398	0.08235	10.2752	0.0013	0.768
Income	1	-0.89088	0.14063	40.1317	<.0001	0.410

Si on regarde les résultats du modèle de régression pour Hilary Clinton on remarque des coefficients ont des signes opposés pour les variables communes avec le modèle de Trump. Mais de plus, on a moins de variables significatives en général. Parmi celles qui ont une influence statistiquement significative au risque de 5% sur les résultats de Hilary, on ne voit pas le niveau de pauvreté et le nombre d'autorisations de construire. Le coefficient d'une variable 'revenue' est devenue positive et significative, par contre le taux d'accession à la propriété maintenant a un lien négatif avec les résultats de Clinton. Le grand pourcentage de personnes nées à l'étranger affecte négativement et au même niveau que dans le 1 classe (nous avons des coefficients autour 0,36) les votes pour Clinton. On peut remarquer que dans cette classe, Clinton a attiré des votes grâce aux jeunes et la population d'origine latino-américaine.

Tableau 4 Clinton Classe 2

Estimations par l'analyse du maximum de vraisemblance						
Paramètre	DDL	Valeur estimée des paramètres	Erreur type	Khi-2	Pr > Khi-2	Rapport de risque
per_point_diff_2012	1	-5.38633	0.21319	638.3316	<.0001	0.005
HSG495213	1	-0.40460	0.17322	5.4561	0.0195	0.667
Edu_batchelors	1	-0.82005	0.12804	41.0225	<.0001	0.440
POP645213	1	-0.35375	0.06439	30.1829	<.0001	0.702
HSG096213	1	-0.19759	0.09826	4.0437	0.0443	0.821
RHI825214	1	1.19693	0.13563	77.8779	<.0001	3.310
AGE295214	1	0.18827	0.07377	6.5144	0.0107	1.207
Black	1	-0.65024	0.06429	102.2896	<.0001	0.522
HSD310213	1	-0.25781	0.08131	10.0523	0.0015	0.773
POP715213	1	-0.16074	0.06423	6.2632	0.0123	0.852
BPS030214	1	-0.31712	0.17021	3.4710	0.0625	0.728
Poverty	1	0.12710	0.08777	2.0972	0.1476	1.136
Income	1	0.33188	0.16001	4.3018	0.0381	1.394

Classe 3

Finalement, on passe à la troisième classe, qui représente une autre cible : citoyens américains faiblement éduqués avec un revenu assez bas. La procédure stepwise nous a permis d'obtenir les modèles suivants :

Trump*= per_point_diff_2012 + HSG495213 + Edu_batchelors + SBO315207 HSG096213+ RHI825214 + Black + HSD310213 +MAN450207 + Poverty + Income +VET605213*

Clinton*= per_point_diff_2012 + HSG445213+ HSG495213 + Edu_batchelors +POP645213 +SBO315207 + RHI825214 + AGE295214 + Black + HSD310213 + MAN450207 + POP715213+ Poverty+ Income + VET605213*

En comparaison avec des classes précédentes, on a une nouvelle variable qui sort : le nombre de vétérans, qui ne caractérise pas cette classe mais qui apparemment joue son rôle en distribution des votes dans ces comptés. On voit ici les variables qui caracterise bien cette classe, comme le nombre de personne par ménage, le pourcentage des noirs-américaines et le nombre de foyers dans les bâtiments comprenant plusieurs unités.

Ensuite, si on parle des modèles, celui de Clinton comporte plus des variables (15 variables) que pour Trump (12 variables), c'est-à-dire il y a des facteurs qui sont susceptible au choix pour le candidat démocrate.

En regardant le test global des modèles on constate que les p-value sont inférieure à 0,05, dons on peut dire qu'au risque de 5% on rejette H0, et qu'il existe au moins une influence d'un des caractéristiques des comptés sur les résultats des votes dans cette classe.

Tableau 5 Trump Classe 3

Estimations par l'analyse du maximum de vraisemblance						
Paramètre	DDL	Valeur estimée des paramètres	Erreur type	Khi-2	Pr > Khi-2	Rapport de risque
per_point_diff_2012	1	5.03554	0.07768	4202.6912	<.0001	153.782
HSG495213	1	0.82057	0.04564	323.1933	<.0001	2.272
Edu_batchelors	1	1.32060	0.04306	940.7469	<.0001	3.746
SBO315207	1	-0.17135	0.06005	8.1424	0.0043	0.843
HSG096213	1	0.23288	0.03396	47.0327	<.0001	1.262
RHI825214	1	-1.03367	0.04937	438.3068	<.0001	0.356
Black	1	0.25656	0.05470	21.9978	<.0001	1.292
HSD310213	1	-0.08176	0.03202	6.5207	0.0107	0.921
MAN450207	1	0.36476	0.06797	28.8015	<.0001	1.440
Poverty	1	-0.19310	0.03964	23.7305	<.0001	0.824
Income	1	-0.49629	0.05325	86.8727	<.0001	0.609
VET605213	1	0.21474	0.05552	14.9622	0.0001	1.240

En examinant les estimations des paramètres du modèle pour Trump, nous nous apercevons que toutes les variables sont significatives au risque de 5%. Si on regarde par exemple sur la variable 'revenue' on constate que le coefficient est moins grand que pour la classe 2, on peut expliquer ça par le fait que Trump cible plutôt les gens avec un faible revenu.

En moyenne, les comtés de la deuxième classe gagnent plus que dans la troisième classe. On a la même situation avec le pourcentage des noirs américaines, en général cette variable est liée positivement avec les votes pour Trump, juste que dans la classe 2 ce lien est plus forte que dans la classe 3. On trouve que c'est bien logique parce que la dernière classe est déjà caractérisée par le pourcentage fort des afro-américaines et ici Trump cible plutôt une pauvre population en général en incluant les derniers, et les comptés 'classique blancs' sont très attirés par la politique de Trump prône pour l'immigration. C'est-à-dire si le pourcentage des gens d'origine augmente dans ces comptés ils sont plus susceptibles de voter pour lui. En ce qui concerne, le pourcentage des blancs, cette variable est très liée négativement avec le

pourcentage de votes pour Trump que dans la classe 2. On a déjà dit que le candidat Republican est intéressé plutôt à les population ‘mélangées’. Mais si on parle de la classe deux, c’est nécessaire dans ces comptés pour le soutien de la politique d’immigration stricte, les polémiques de la demande d’asile syrienne, de la loi contre l’avortement, du système de santé, de la tension religieuse (islam) que dans les comtés de la classe 3, Trump plutôt cherche la population mal éduquée et pauvre, qui ici représenté par les personnes d’origine étrangére. De plus dans cette classe on voit les vétérans, mais qui sont plutôt assez pauvres, mais gardent quand même leur patriotisme et influencent positivement le pourcentage des votes pour Trump, qui prône la politique de protectionnisme et l’idée d’unicité des Etats-Unis avec son devise ‘Rendre sa grandeur à l'Amérique’.

Tableau 6 Clinton Classe 3

Estimations par l'analyse du maximum de vraisemblance						
Paramètre	DDL	Valeur estimée des paramètres	Erreur type	Khi-2	Pr > Khi-2	Rapport de risque
per_point_diff_2012	1	-4.63945	0.07071	4305.2542	<.0001	0.010
HSG445213	1	0.15829	0.03990	15.7350	<.0001	1.172
HSG495213	1	-0.56152	0.04991	126.5655	<.0001	0.570
Edu_batchelors	1	-1.06985	0.04782	500.6026	<.0001	0.343
POP645213	1	-0.13096	0.06117	4.5839	0.0323	0.877
SBO315207	1	0.11488	0.05204	4.8736	0.0273	1.122
RHI825214	1	0.54153	0.05849	85.7142	<.0001	1.719
AGE295214	1	0.10620	0.03810	7.7679	0.0053	1.112
Black	1	-0.82665	0.05945	193.3431	<.0001	0.438
HSD310213	1	-0.14004	0.04369	10.2743	0.0013	0.869
MAN450207	1	-0.22802	0.07947	8.2325	0.0041	0.796
POP715213	1	-0.06728	0.03275	4.2211	0.0399	0.935
Poverty	1	-0.29946	0.04331	47.8006	<.0001	0.741
Income	1	0.17981	0.06021	8.9189	0.0028	1.197
VET605213	1	-0.29103	0.07318	15.8152	<.0001	0.747

En examinant les estimations des paramètres du modèle pour Hilary Clinton nous remarqons également que toutes les variables sont significatives au risque de 5%. Pour Clinton on voit aussi que des coefficients pour les variables communes avec le modèles de Trump ont des signes opposés, ce qui est bien logique. On est plutôt intéressé par les variables qui sont particulières dans ces comptés pour ce candidat et qui peuvent attirer des votes de Trump. Ce sont les pourcentages des jeunes et le taux d'accession à la propriété. En ce qui concerne la première variable on peut constater que son lien avec le pourcentage des votes pour Clinton est plus fort dans la deuxième classe. En effet, Hilary a le support des jeunes américaines qui est plutôt progressif dans les comtés avec une population plutôt ‘blanche’ et un niveau de revenu moyen ou plus. Par contre, le taux d'accession à la propriété joue un rôle positif dans la classe 3 en comparaison avec la classe 2. Dans ces comptés cet indicateur correspond aux gens au-dessus du niveau de pauvreté, que dans les comtés de la classe 2, il représente des américains blancs typiques.

Conclusion :

En guise de conclusion, les votes des élections américaines de 2016 entre les deux candidats révèlent plusieurs préférences selon les caractéristiques économiques des comtés dans chacun des 50 états d'Amérique.

Notre étude a été effectuée en commençant par des statistiques descriptives qui pour structurer et représenter l'information contenue dans notre base.
La population qui est l'ensemble des sujets observés : les comtés.
Le caractère est la propriété étudiée sur ces sujets : les indicateurs économique des états unis.

L'ACP nous a permis d'avoir des représentation et des informations traitées sur nos 3112 individus et 34 variables, en ressortant des données brutes et les éventuels liens existant entre les variables (en terme de corrélation), afin de découvrir quelles sont les tendances dominantes de l'ensemble de données, et réduire efficacement le nombre de dimensions étudiées (et ainsi simplifier l'analyse), en cherchant à exprimer le plus fidèlement possible l'ensemble original de données grâce aux relations détectées entre les variables.

Parmi les objectifs de la régression on peut identifier le fait de décrire le processus de causalité entre les variables. A travers ça, on étudie le signe et la valeur des coefficients en conceptualisant le rôle des différentes variables dans l'explication des valeurs prises par Y. Si une variable est redondante, elle doit être ventilée au risque de perturber et rendre illusoire la lecture les résultats. La problématique alors est de définir une stratégie de détection de multi colinéarité avant toute interprétation approfondie.

Pour répondre à cette problématique, il existe plusieurs pistes et sans remettre en doute la pertinence de toutes ces dernières, nous avons limiter notre choix a deux méthodes. Premièrement, le calcul des corrélations croisées entre chaque variable en utilisant la méthode de Spearman pour sa simplicité même dans les situations multiples. Deuxièmement, pour ajouter de la robustesse à nos résultats, on a utilisé le facteur d'inflation de la variance qui est par exemple contrairement au teste de Klein qui ne détecte que la colinéarité variée, le VIF évalue la multi colinéarité.

Il est à noter qu'il faut toujours avoir un point de vue critique coté métier pour appréhender les interdépendances en jeu.

La classification nous a permis d'avoir des groupes de comtés homogènes, pour former des unités systématiques hiérarchisées en fonction de différents ressorts sociaux économiques.

En réalisant la régression sur les 3 classes de comtés on a bien remarqué que dans chaque classe il existe un lien particulier entre les résultats d'élection et les caractéristiques des comptés. En général, on peut dire que Hilary gagne les comptés avec une population et une densité élevée, qui représente des américains bien éduqués avec un haut revenu, mais qui ne sont pas nombreux au final. Par contre Trump a ciblé les comptés avec une proportion important de blancs de moyenne classes ou avec une proportion importante de Noirs d'Amérique de classe plus basse que la moyenne.