Étude de marché

Objet : exportation de poulet à l’international

Enjeux

- Cibler des pays selon :
  - **Leur besoin en alimentation**
  - **Leur pouvoir d’achat**

Données

- Étude basée sur **5** variables :

  - **Croissance démographique**
  - **Disponibilité alimentaire par habitant**
  - **Disponibilité alimentaire en protéines par habitant**
  - **Proportion de protéines d'origine animale**
  - **PIB par habitant**

- Données alimentaires issues du site de la **FAO**
- PIB par habitant issu du site **banquemondiale.org** (sources supplémentaires citées en fin de présentation)
- Chiffres : **année 2017**

Objectif

- Classer les pays en **5 groupes** selon les variables
- Identifier le groupe pour lequel :

    - Croissance démographique **forte**
    - Disponibilité alimentaire par habitant **faible**
    - Disponibilité alimentaire en protéines par habitant **faible**
    - Proportion de protéines d'origine animale **faible**
    - PIB par habitant **fort**

Démarche

- Analyse en Composantes Principales : **ACP**
- Identification des **variables synthétiques**
- Partitionnement ou **clustering**
- **Identification des clusters** selon variables synthétiques
- Identification des **pays cibles**

Analyse en composantes principales (ACP)

Définition

- Permet de **regrouper** des **variables très corrélées** en une **variable synthétique**
- Simplifie l'analyse multidimensionnelle, impossible à représenter géométriquement au-delà de trois dimensions
- Principe : projection orthogonale sur le premier axe principal d'inertie, puis si besoin sur le second axe, orthogonal au premier.
- Plusieurs axes d'inertie peuvent être utilisés. Nombre déterminé par l'éboulis des valeurs propres

Contrainte

Les variables ont des unités différentes : **comment regrouper les données en une seule variable sans générer d'erreurs ?**

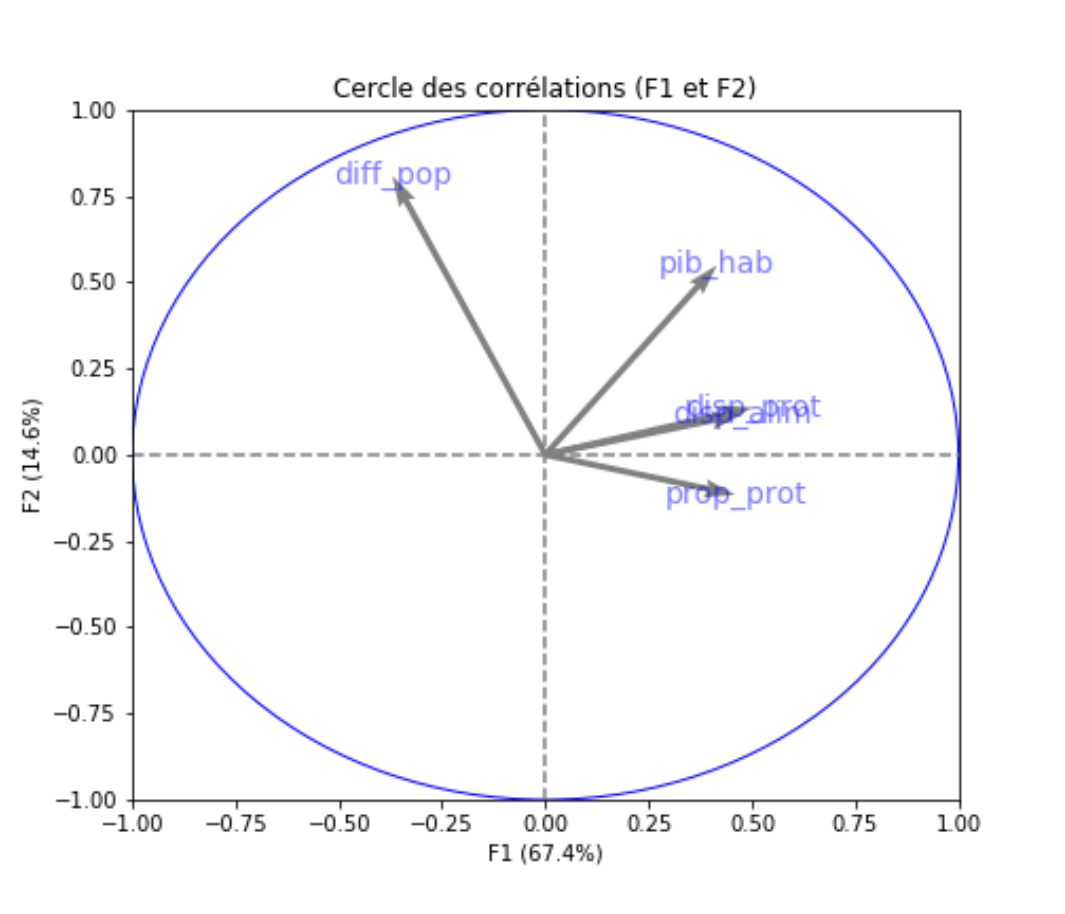
Solution : centrage et réduction

- **Centrage** : égalité des **moyennes** en retranchant la moyenne d'une population à chaque individu.
- **La moyenne est donc égale à 0**

- **Réduction** : égalité des **variances** en divisant les données centrées par leur écart-type.
- **La variance est ainsi égale à 1**

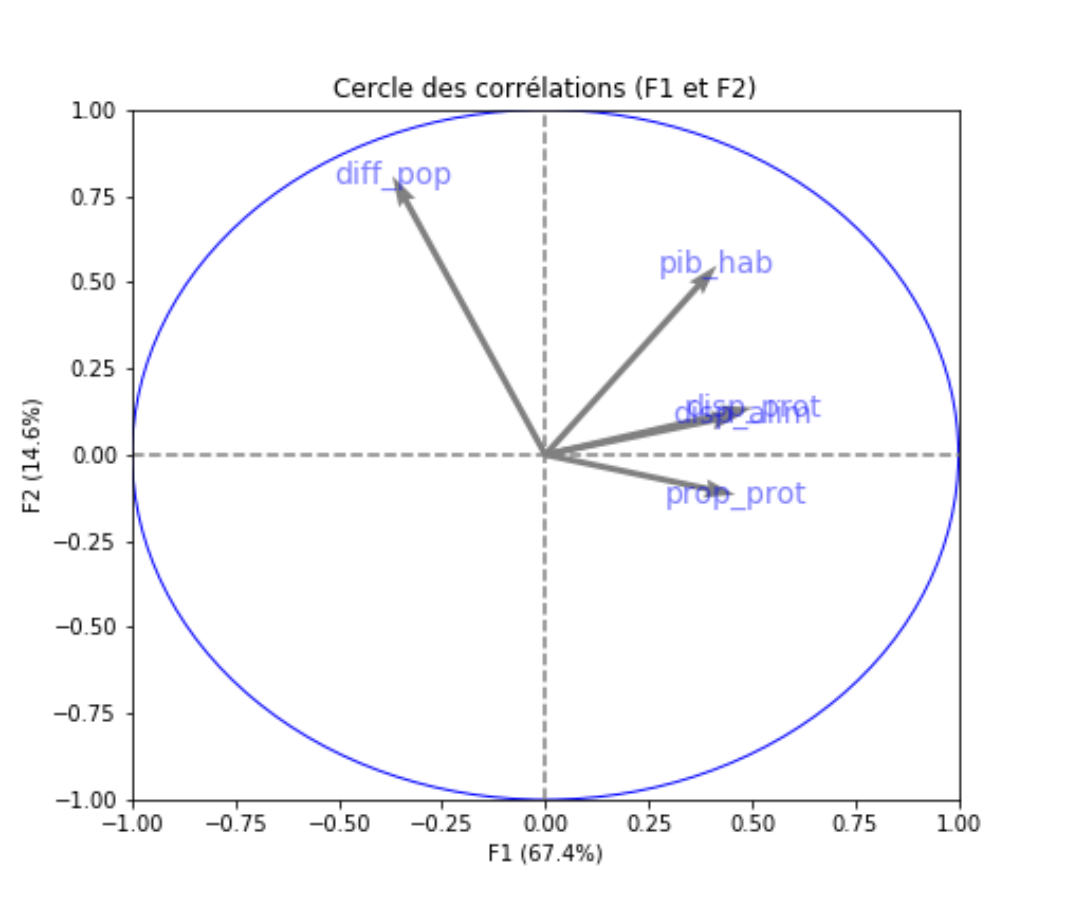
$$X_{cr} = \frac{X - \bar{X}}{s_X}$$

Cercle des corrélations



- F1 et F2 sont **les principaux axes d'inertie**
- Ils forment le **premier plan factoriel**
- F1 et F2 sont également les **variables synthétiques**

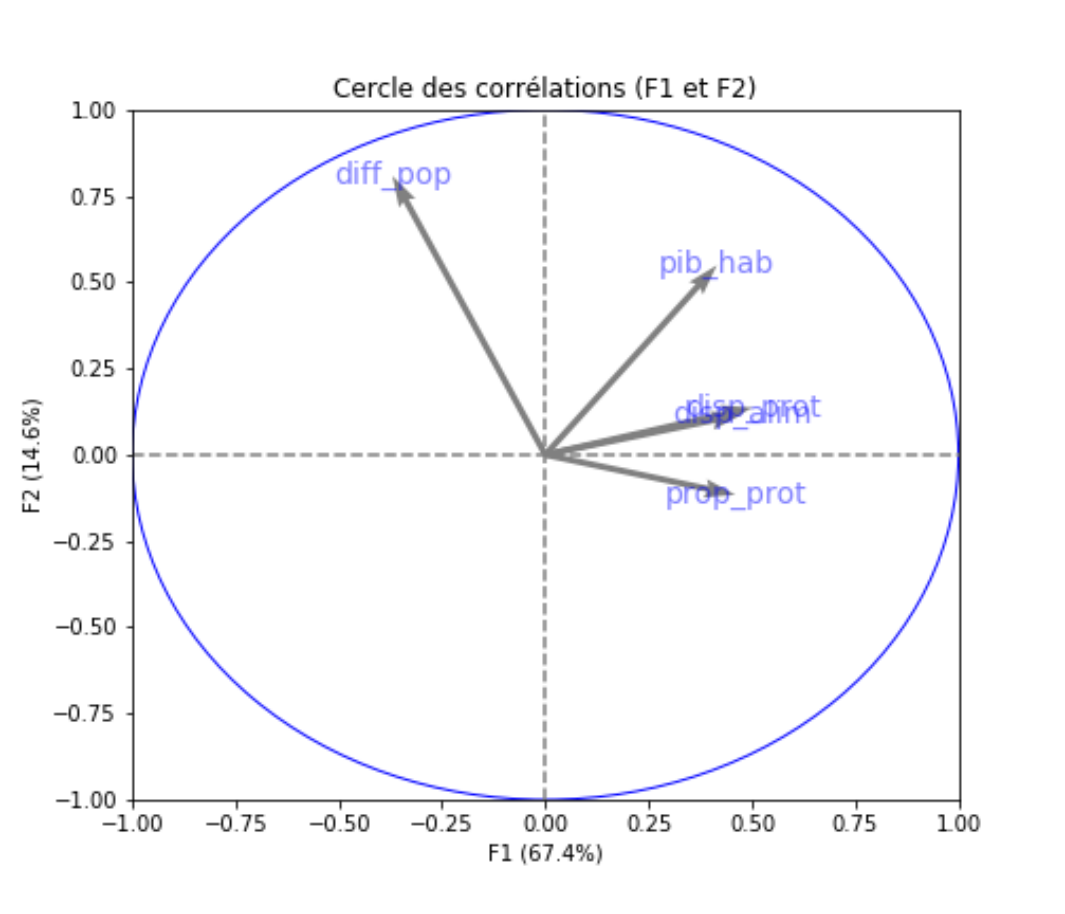
Interprétation



- Les pourcentages indiquent que F1 et F2 représentent plus de **80% de l'inertie totale**

- Les flèches sont assez longues : **les variables sont donc bien représentées sur ce plan**

Interprétation



- F1 est corrélée à:
  - disponibilité alimentaire totale
  - disponibilité protéines
  - proportion protéines

- F2 est corrélée à :
  - Croissance
  - PIB/habitant

Conclusion ACP

- 2 variables synthétiques F1 et F2 :
    - F1 : variable de l'**accès à la nourriture**
    - F2 : variable **socio-économique**
- Objectif : trouver les pays dont l'alimentation est faible en protéines et avec un retour sur investissement assuré, soit :
    - **F1 faible**
    - **F2 élevée**

Partitionnement (ou clustering)

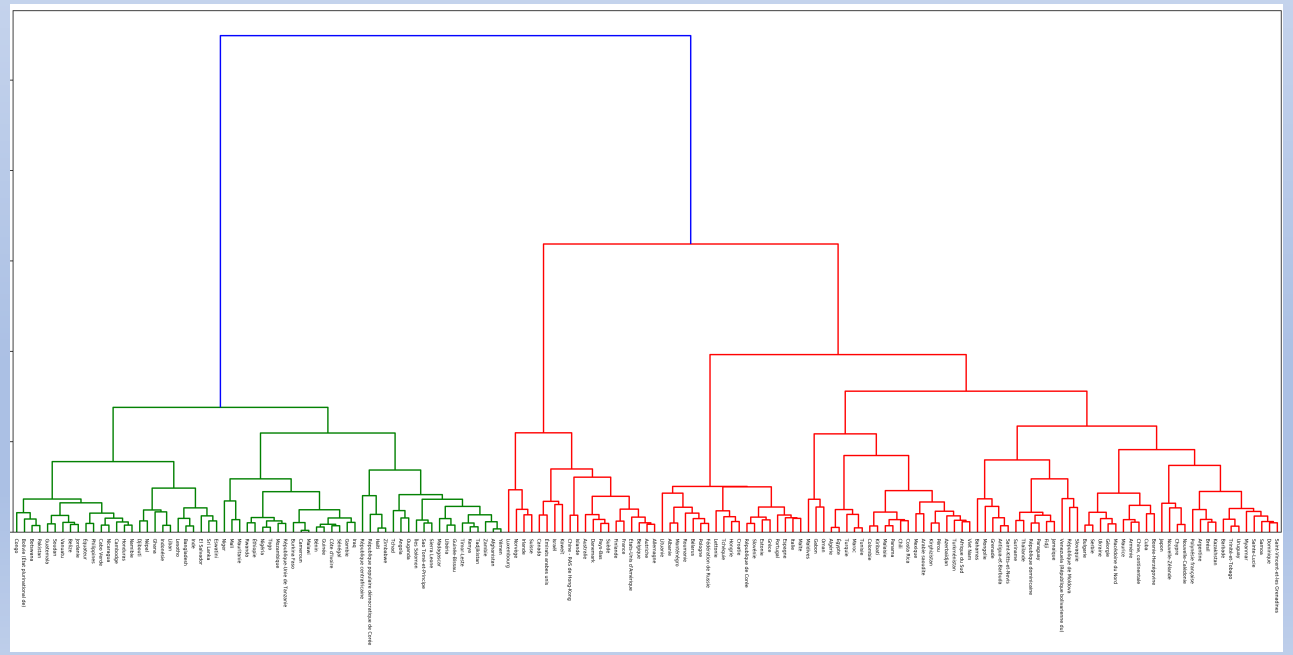
2 Méthodes de partitionnement

Classification hiérarchique

Algorithme du K-Means

Définition classification hiérarchique

- Division du nuage de points selon leur resserrement sur eux-mêmes, c'est la notion d'**inertie intraclasse**
- Résultat affiché sous forme d'un **dendrogramme**



Définition algorithme du K-Means

- Détermine les clusters selon le placement de **centres de gravité, ou centroïdes**, pour chacun de ces clusters.
- Les clusters n'étant justement pas connus, les centres de gravité sont donc placés en associant **aléatoirement** les points dont ils découlent.
- Les centroïdes se repositionnent à chaque itération en associant à chaque fois de nouveaux individus jusqu'à tendre vers un **équilibre**
- Lorsque l'algorithme **converge**, les clusters sont déterminés.

Résultat cluster 1 classification : 64 pays

Afghanistan	Cabo Verde	Djibouti	Indonésie	Lesotho	Népal	Timor-Leste	Tanzanie
Angola	Rép. Centrafricaine	Gambie	Iraq	Libéria	Vanuatu	Zimbabwe	Togo
Bangladesh	Sri Lanka	Ghana	Côte d'Ivoire	Madagascar	Nicaragua	Rwanda	Ouganda
Bolivie	Tchad	Guatemala	Jordanie	Malawi	Niger	Sao Tomé-et-Principe	Burkina Faso
Botswana	Congo	Guinée	Kenya	Mali	Nigéria	Sénégal	Éthiopie
Belize	Bénin	Haïti	Cambodge	Mauritanie	Pakistan	Sierra Leone	Yémen
Îles Salomon	Équateur	Honduras	Corée du Nord	Mozambique	Philippines	Tadjikistan	Zambie
Cameroun	El Salvador	Inde	Liban	Namibie	Guinée-Bissau	Eswatini	Soudan

Résultat cluster 1 K-Means : 56 pays

Afghanistan	Zambie	Djibouti	Ouganda	Lesotho	Népal	Timor-Leste
Angola	Rép. Centrafricaine	Gambie	Iraq	Libéria	Vanuatu	Zimbabwe
Bangladesh	Yémen	Ghana	Côte d'Ivoire	Madagascar	Nicaragua	Rwanda
Soudan	Tchad	Guatemala	Jordanie	Malawi	Niger	Sao Tomé-et-Principe
Botswana	Congo	Guinée	Kenya	Maldives	Nigéria	Sénégal
Belize	Bénin	Haïti	Cambodge	Togo	Pakistan	Sierra Leone
Îles Salomon	Éthiopie	Honduras	Corée du Nord	Mozambique	Tanzanie	Tadjikistan
Cameroun	Burkina Faso	Inde	Liban	Namibie	Guinée-Bissau	Eswatini

Résultat cluster 2 classification : 20 pays

Australie	Finlande	Islande	Pays-Bas	Émirats arabes unis
Autriche	France	Irlande	Norvège	USA
Canada	Allemagne	Israël	Suède	Belgique
Danemark	Hong-Kong	Koweït	Suisse	Luxembourg

Résultat cluster 2 K-Means : 19 pays

Australie	Finlande	Islande	Pays-Bas	Émirats arabes unis
Autriche	France	Irlande	Norvège	USA
Canada	Allemagne	Israël	Suède	Belgique
Danemark	Hong-Kong		Suisse	Luxembourg

Résultat cluster 3 classification : 19 pays

Albanie	Croatie	Malte	Russie
Belarus	Italie	Tchéquie	Slovénie
Estonie	Corée du Sud	Pologne	Espagne
Grèce	Lettonie	Portugal	Monténégro
Hongrie	Lituanie	Roumanie	

Résultat cluster 3 K-Means : 36 pays

Albanie	Croatie	Malte	Russie
Belarus	Italie	Tchéquie	Slovénie
Estonie	Corée du Sud	Pologne	Espagne
Grèce	Lettonie	Portugal	Monténégro
Hongrie	Lituanie	Roumanie	Samoa
Arménie	Barbade	Bulgarie	Cuba
Argentine	Brésil	Chine	Polynésie française
Bosnie-Herzégovine	Kazakhstan	Japon	Maurice
Nouvelle-Zélande	Trinité-et-Tobago	Ukraine	Uruguay

Résultat cluster 4 classification : 22 pays

Algérie	Égypte	Maldives	Arabie saoudite	Turquie
Chili	Gabon	Mexique	Afrique du Sud	Viet Nam
Colombie	Kiribati	Maroc	Turkménistan	
Costa Rica	Kirghizistan	Panama	Oman	
Azerbaïdjan	Malaisie	Pérou	Tunisie	

Résultat cluster 4 K-Means : 19 pays

Algérie	Égypte		Arabie saoudite	Turquie
Chili	Gabon	Mexique	Afrique du Sud	Viet Nam
		Maroc	Turkménistan	Koweït
	Kirghizistan		Oman	Mali
Azerbaïdjan		Pérou	Tunisie	Mauritanie

Résultat cluster 5 classification : 40 pays

Arménie	Chine	Bosnie-Herzégovine	Nouvelle-Calédonie	Suriname
Antigua-et-Barbuda	Cuba	Grenade	Macédoine du Nord	Thaïlande
Argentine	Chypre	Kazakhstan	Nouvelle-Zélande	Trinité-et-Tobago
Bahamas	Dominique	Jamaïque	Paraguay	Ukraine
Barbade	Rép. Dominicaine	Japon	Saint-Kitts-et-Nevis	Uruguay
Brésil	Fidji	Maurice	Sainte-Lucie	Venezuela
Bulgarie	Polynésie française	Mongolie	Saint-Vincent-et-les Grenadines	Samoa
Myanmar	Géorgie	Moldavie	Slovaquie	Serbie

Résultat cluster 5 K-Means : 34 pays

Antigua-et-Barbuda	Costa Rica	Géorgie	Moldavie	Saint-Kitts-et-Nevis
Bahamas	Chypre	Kiribati	Nouvelle-Calédonie	Sainte-Lucie
Bolivie	Dominique	Grenade	Macédoine du Nord	Slovaquie
Myanmar	Rép. Dominicaine	Indonésie	Panama	Suriname
Cabo Verde	Équateur	Jamaïque	Paraguay	Thaïlande
Sri Lanka	El Salvador	Malaisie	Philippines	Venezuela
Colombie	Fidji	Mongolie	Serbie	

Comparaison des clusters de chaque algorithme

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Pays en commun	55	19	19	16	20
Part dans Classification hiérarchique	86%	100%	100%	73%	50%
Part dans K-Means	98%	95%	53%	84%	59%

Le cluster 1 représente les plus pays **pauvres**.
Le cluster 2 regroupe les pays les **plus riches**.

Moyenne des composantes principales de chaque cluster classification hiérarchique

Cluster	Moyennes F1	Moyennes F2
1	**-1.885886**	**0.175614**
2	2.935202	1.223856
3	2.089808	-0.708845
4	0.050083	0.311895
5	0.529611	-0.727752

Le seul cluster avec une moyenne de F1 négative et une moyenne de F2 positive est le **cluster 1** : ce cluster regroupe nos pays-cible.

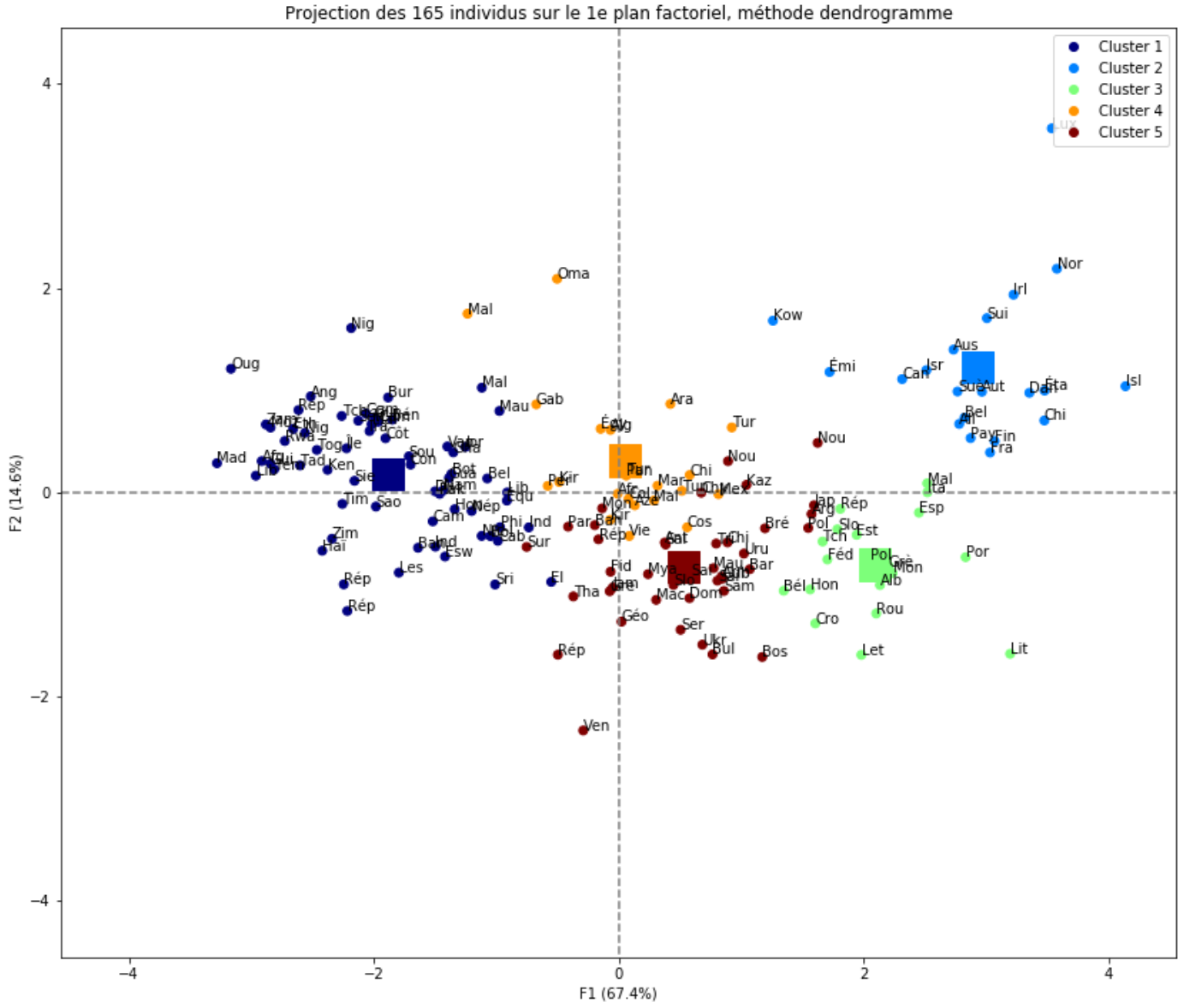
Moyenne des composantes principales de chaque cluster du K-Means

Cluster	Moyennes F1	Moyennes F2
1	**-2.029224**	**0.260956**
2	3.023284	1.223856
3	1.590380	-0.681837
4	0.030456	0.175614
5	-0.094958	-0.628989

Là aussi, un seul cluster affiche une moyenne de F1 négative et une moyenne de F2 positive : le **cluster 1** .

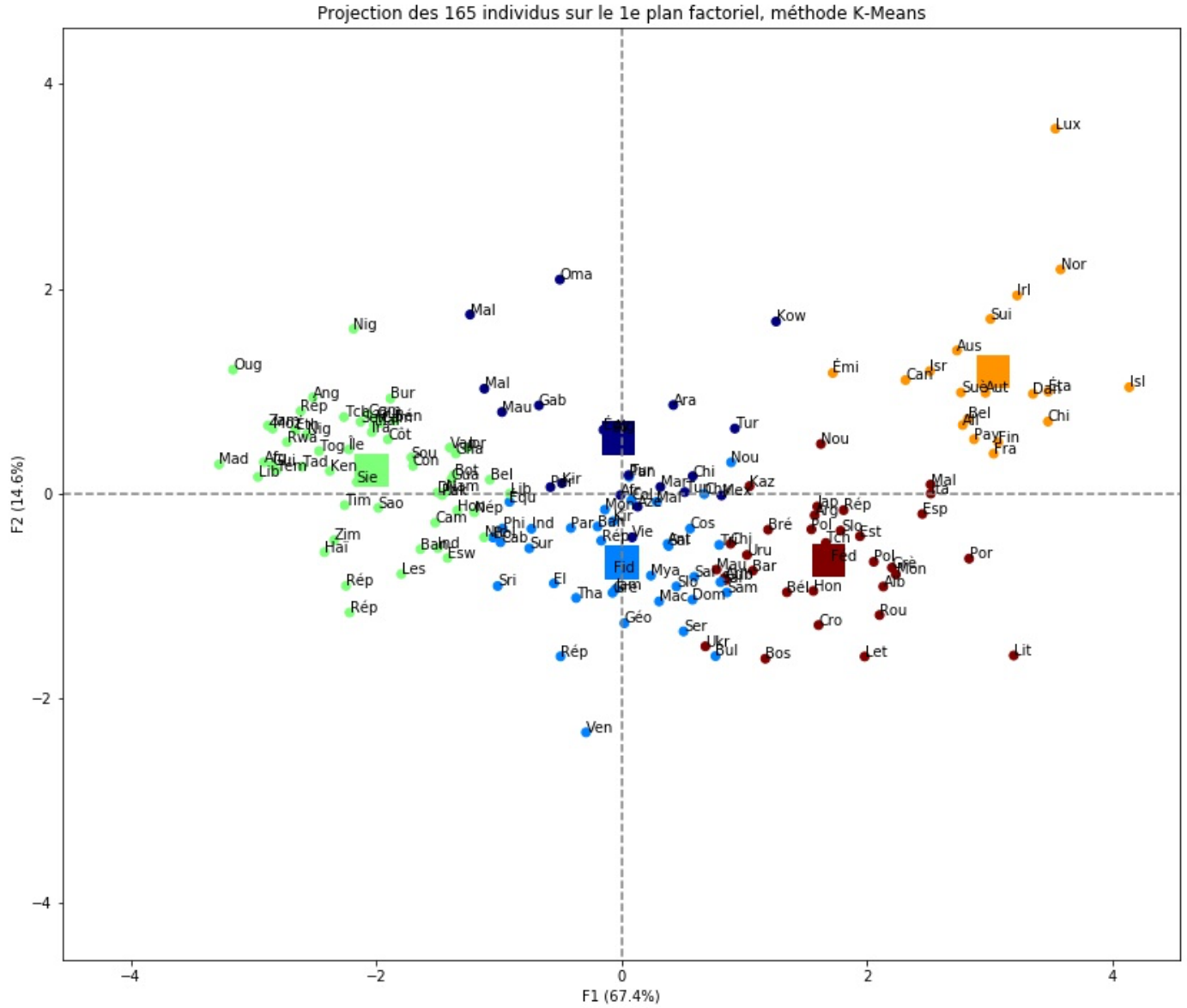
Projection des individus : Classification hiérarchique

En observant les centroïdes (carrés) et leur cluster associé, nous déduisons que le **cluster 1** rassemble nos **pays cibles**.



Projection des individus : K-Means

Ici, c’est le **cluster vert** qui rassemble nos **pays cibles**.



Pays cibles : Classification ou K-Means ?

Approche économique

- Au sein du cluster retenu pour chaque algorithme, il y a logiquement les pays décimés par la **famine** (exemple : le Yémen)
- Ces pays doivent rester la cible **d'actions humanitaires** et non pas commerciales
- Il convient donc de **retirer ces pays des clusters retenus**

Démarche

- Pour le cluster 1 de chaque algorithme :
    - Suppression des pays sous le **seuil de pauvreté** (moins de 1,9$ US/jour par habitants)
    - Sélection des **10 pays avec F1 le plus bas**
    - Sélection des **10 pays avec F2 le plus haut**
    - **Intersection** de ces classements
    - **Comparaison** des résultats pour le cluster de chaque algorithme

Résultats et caractéristiques

	Zone	diff_pop	disp_alim	disp_prot	pib_hab	prop_prot
0	Ouganda	3.827117	782560.0	19107.75	910.269755	20.783190
1	République-Unie de Tanzanie	3.037009	874540.0	21819.70	937.301492	15.623954
2	Angola	3.377933	828550.0	19750.15	3409.929285	30.456478

Conclusion

Pays cibles

Ouganda
Tanzanie
Angola

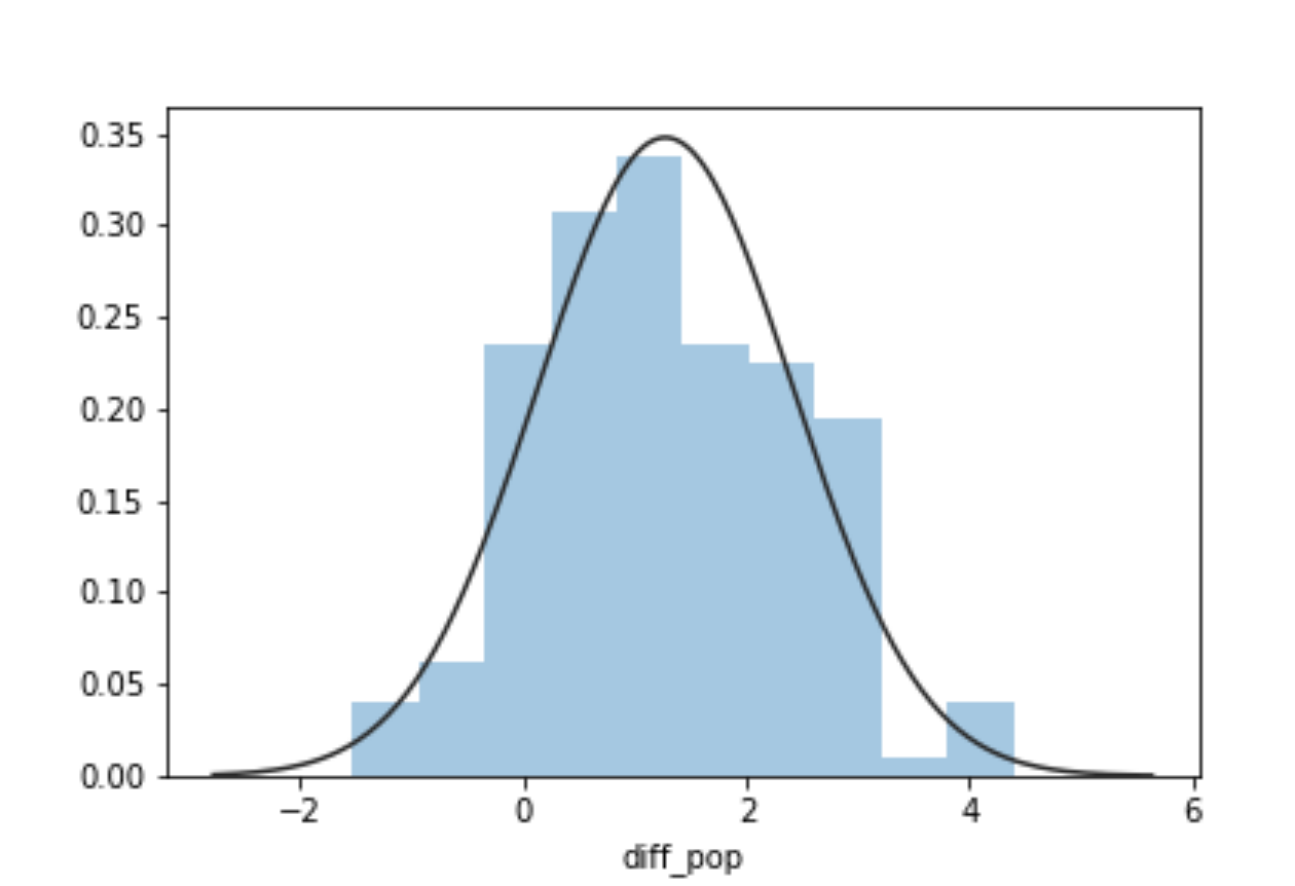
Sources données PIB

- PIB par habitant : https://donnees.banquemondiale.org/
- Elle est complétée par les sources suivantes pour ces pays :
- - République démocratique populaire de Corée : https://www.populationdata.net/pays/coree-du-nord/
- - Djibouti : https://perspective.usherbrooke.ca/bilan/servlet/BMTendanceStatPays?langue=fr&codePays=DJI&codeStat=NY.GDP.PCAP.PP.CD&codeStat2=x
- - Érythrée : https://fr.countryeconomy.com/gouvernement/pib/erythree
- - Nouvelle-Calédonie : https://www.lefigaro.fr/economie/le-scan-eco/2017/12/02/29001-20171202ARTFIG00004-du-nickel-au-tourisme-8-choses-a-savoir-sur-l-economie-de-la-nouvelle-caledonie.php
- - Polynésie française : http://www.polynesie-francaise.pref.gouv.fr/content/download/28076/146842/file/EC_Regards-eco-PF_oct2017.pdf (PIB 2016)
- - Venezuela : https://fr.wikipedia.org/wiki/%C3%89conomie_du_Venezuela
- Seuil de pauvreté : https://www.banquemondiale.org/fr/research/brief/poverty-and-shared-prosperity-2018-piecing-together-the-poverty-puzzle-frequently-asked-questions

Annexes

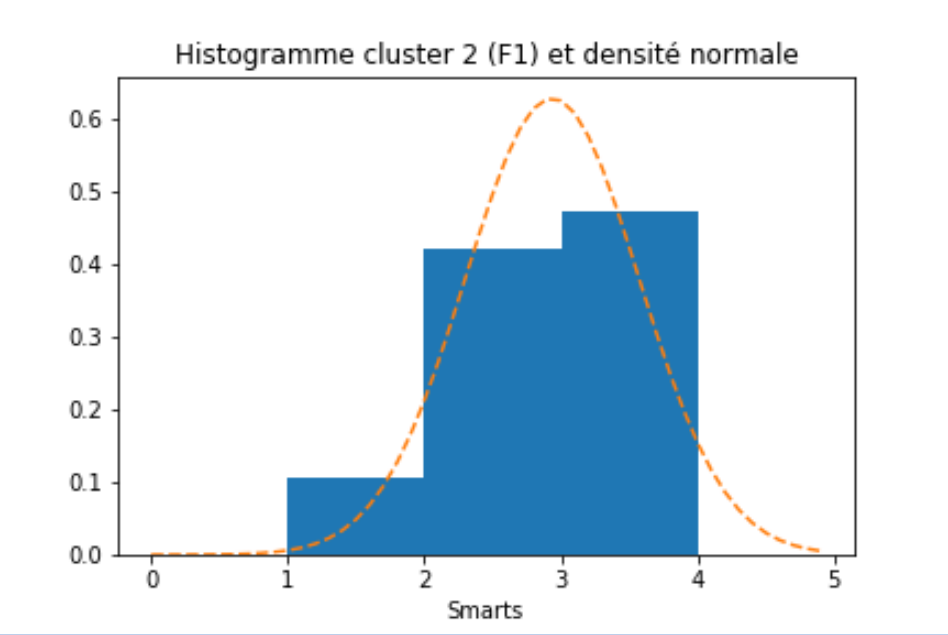
Test d’adéquation

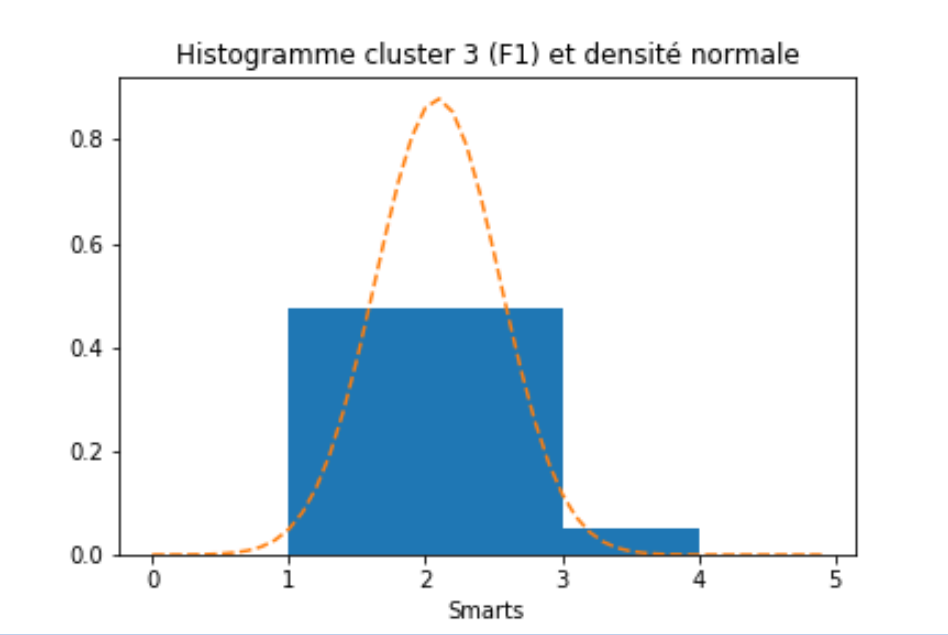
Parmi les variables utilisées, la **différence de population** entre 2016 et 2017 (en %) suit la **loi normale**.



Test de comparaison

- Test effectué sur les **clusters 2 et 3 de la classification hiérarchique**
- La variable F1 des deux clusters est **gaussienne**





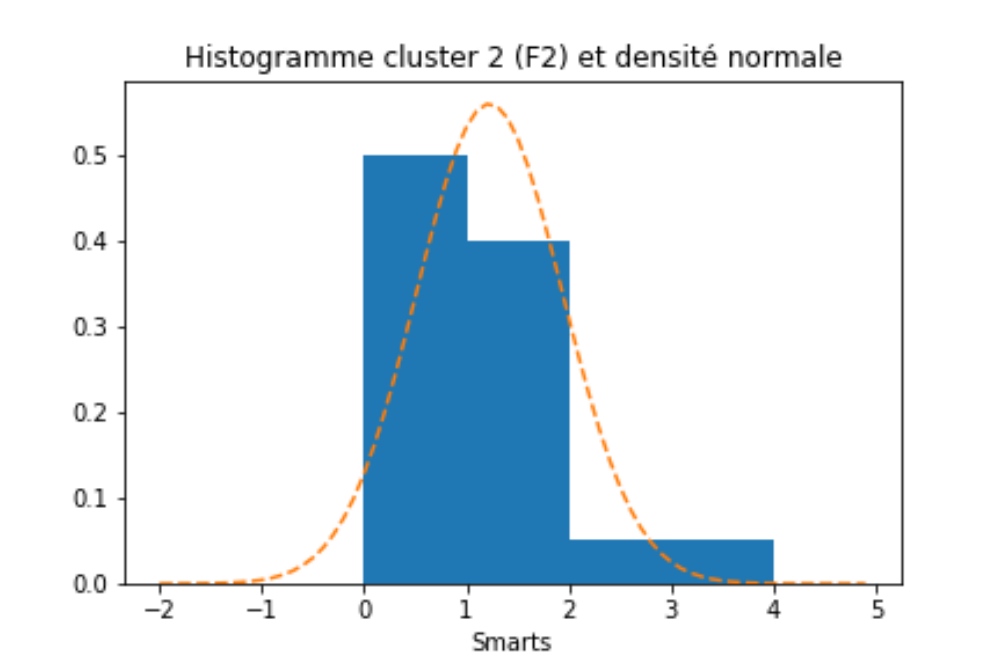
Test de comparaison

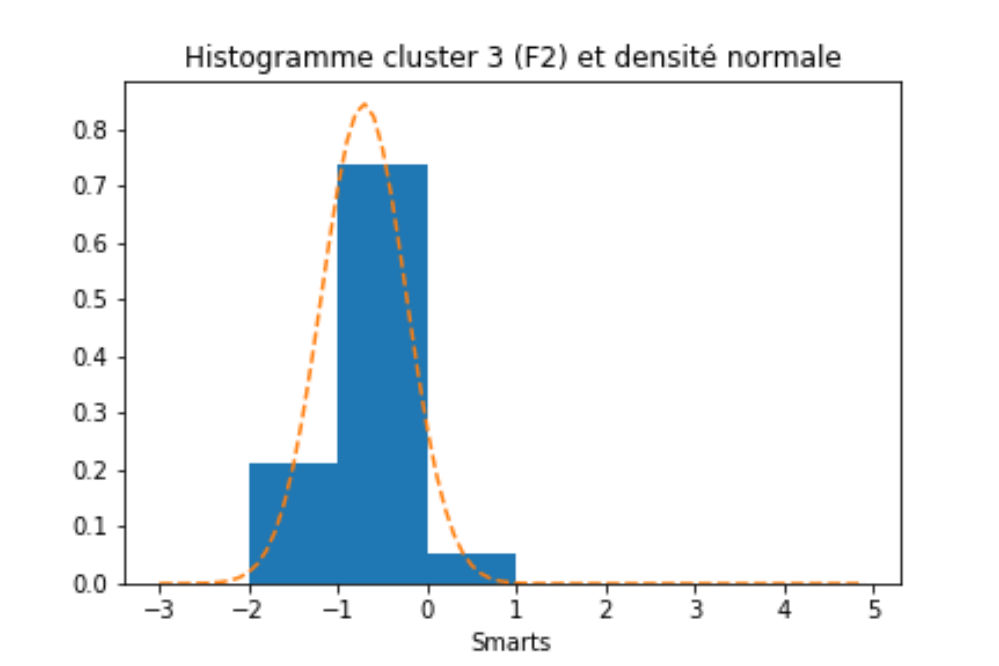
- Égalité des variances de F1 :
  - p-value = 0.16
  - Elle est est supérieure au niveau de test 5%
  - **On ne rejette pas l'égalité des variances**

- Égalité des moyennes de F1 :
  - p-value = 4.28e-05
  - Elle est est inférieure au niveau de test 5%
  - **On peut rejeter l'hypothèse d'égalité des moyennes**

Test de comparaison

- La variable F2 des deux clusters est **gaussienne**





Test de comparaison

- Égalité des variances de F2 :
  - p-value = 0.09
  - Elle est est supérieure au niveau de test 5%
  - **On ne rejette pas l'égalité des variances**

- Égalité des moyennes de F2 :
  - p-value = 1.15e-11
  - Elle est est inférieure au niveau de test 5%
  - **On peut rejeter l'hypothèse d'égalité des moyennes**

Test de comparaison

- L’hypothèse d’égalité des moyennes est rejetée pour les deux dimensions F1 et F2

- **Les clusters 2 et 3 ne suivent donc pas la même distribution.**