CNAM STA101

Analyse des données: méthodes descriptives

2019-2020 session 2

_____

François LABASTIE - Olga KOROLEVA

_____

PROJET

Consommations et habitudes alimentaires en France métropolitaine

Introduction

Ce projet a pour objectif d'observer une tranche de la population française en fonction des habitudes alimentaires et facteurs d'activité physique ou sédentarité, et d'explorer les éléments pouvant conduire au surpoids et à l'obésité.

Nous travaillerons sur les données issues de l'étude *Individuelle Nationale des Consommations Alimentaires*[1] - INCA3 - de l'ANSES conduite en 2014-2015 et publiée en open data en décembre 2019 sur le site data.gouv.fr. L'ANSES est l'*agence nationale de sécurité sanitaire, de l'alimentation, de l'environnement et du travail.*

INCA3 est une enquête transversale visant à estimer les consommations alimentaires et les comportements en matière d'alimentation des individus vivant en France métropolitaine (hors Corse).

Les données recueillies dans l'étude portent sur diverses thématiques en lien avec l'évaluation des risques nutritionnels ou sanitaires liés à l'alimentation : consommations d'aliments et boissons ainsi que les habitudes alimentaires (occasions et lieux de consommation, etc.). Des données sur l'activité physique et niveau de sédentarité ainsi que des informations sociodémographiques et anthropométriques ont également été recueillies.

Obésité et surpoids : les dangers pour la santé

L'étude INCA3 cherche à évaluer les risques sanitaires en vue d'éclairer la décision publique et assurer la sécurité des consommateurs. A notre niveau, nous choisirons d'orienter notre projet d'exploration des donnés selon les risques et comportements liés au surpoids et à l'obésité.

Selon l'Organisation Mondiale de la Santé, le nombre de cas d'obésité a presque triplé à l'échelle mondiale entre 1975 et 2016. Et l'OMS résume ainsi les observations sur ces critères [2] :

- Obésité → 13% de la population adulte mondiale (11% hommes - 15% femmes).
- Surpoids → 39% des adultes (39% hommes et 40% femmes).

Le principe de base est qu'une personne ayant un surplus de poids serait plus à risque de développer divers problèmes de santé (cholestérol, diabète, hypertension et maladies cardiovasculaires).

1 https://www.data.gouv.fr/fr/datasets/donnees-de-consommations-et-habitudes-alimentaires-de-letude-inca-3/
Anses. (2017). Rapport de l'Anses relatif à la troisième étude individuelle nationale des consommations alimentaires (Etude INCA3). Actualisation de la base de données des consommations alimentaires et de l'estimation des apports nutritionnels des individus vivant en France. (saisine 2014-SA-0234). Maisons-Alfort : Anses, 535 p. https://www.anses.fr/fr/system/files/NUT2014SA0234Ra.pdf

2 https://www.who.int/fr/news-room/fact-sheets/detail/obesity-and-overweight

Échelle d'observation

L'observation de l'obésité se base sur l'IMC - indice de masse corporelle - qui est une mesure simple du poids par rapport à la taille [3] :

- IMC = poids / (taille*taille) [unité → kg/m2]

Une classification des obésités a été établie par l'Organisation mondiale de la santé :

- IMC < 18,5 → caractérise la maigreur
- 25<IMC<30 → simple "surpoids"
- 30<IMC<35 → obésité dite modérée
- 35<IMC<40 → obésité considérée sévère
- 40<IMC → obésité morbide

Pour notre projet, l'IMC est un indicateur de référence comme aide à l'analyse.

Objectifs de notre étude

Deux questions surgissent dans notre exploration des données :

- Quelles sont les pratiques alimentaires de notre échantillon ?
- Quels sont les groupes caractéristiques dans la population ?

Nous y répondrons à l'aide d'analyses univariées et bivariées de nos données, suivies d'analyses en composantes principales - ACP - afin de condenser l'information contenue dans notre tableau de données.

L'analyse des corrélations linéaires entre les variables et la visualisation graphique des distances entre individus nous permettra de dégager les liaisons entre variables et les ressemblances entre individus.

Nous chercherons ensuite à créer des groupes d'individus en partitionnant les données à l'aide des algorithmes CAH - Classification Ascendante Hiérarchique - et K-means.

Outils techniques

Pour la réalisation de ce projet, nous employons les technologies suivantes :

- Langage de programmation Python
- Analyse des données avec librairies Numpy & Pandas
- Méthodes factorielles avec librairie Scikit learn
- Visualisation des données avec Matplotlib et Seaborn
- Environnement de développement Anaconda / Jupyter Notebook
- Outil de versionning git / github

3https://www.universalis.fr/encyclopedie/obesite/1-qu-est-ce-que-l-obesite/

A. Données

1. Origine & Description de l'échantillon

L'étude INCA3 est une enquête transversale visant à estimer les consommations alimentaires et les comportements en matière d'alimentation des individus vivant en France. Elle fut menée entre février 2014 et septembre 2015 en 5 vagues d'enquête auprès d'un échantillon représentatif de 5 855 individus vivant en France métropolitaine (hors Corse).

Les individus ont été sélectionnés selon un plan de sondage aléatoire à trois degrés (unités géographiques, logements puis individus), à partir du recensement annuel de la population de 2011, en respectant une stratification géographique (région, taille d'agglomération) afin d'assurer la représentativité sur l'ensemble du territoire.

A partir de cet échantillon d'un total de 5 855 individus, deux échantillons indépendants ont été constitués par l'ANSES, un groupe « Enfants » et un groupe « Adultes »:

- 2 698 enfants (0 - 17 ans)
- 3 157 adultes (18 - 79 ans)

Recueil des données

Les différentes techniques de recueil de données employées dans l'étude INCA3 comprennent des interviews et questionnaires, associés à des matériels référentiels relatifs aux aliments (cahiers de photographies, etc.), de même qu'une visite à domicile. Ainsi, divers types de questionnaires furent employés:

- Un questionnaire administré en face-à-face (FAF)
- Un questionnaire auto-administré (AA)
- Un questionnaire administré par téléphone (TEL)

Les mesures anthropométriques telles que le poids (en kg) et la taille (en cm) des individus ont été mesurés par l'enquêteur lors de la visite à domicile.

Les données recueillies dans l'étude portent sur diverses thématiques en lien avec l'évaluation des risques nutritionnels ou sanitaires liés à l'alimentation : consommations d'aliments et boissons, et habitudes alimentaires.

Des données sur les pratiques d'activité physique et de niveau de sédentarité ainsi que sur les caractéristiques sociodémographiques, anthropométriques et de niveau de vie ont également été recueillies.

2. Sélection des individus

L'échantillon de l'étude INCA3 est composé de 5 855 individus divisé en 2 groupes: adultes et enfants. Mais les auteurs de l'étude INCA3 posent la contrainte suivante :

> *Compte tenu du plan d'échantillonnage retenu pour l'étude, les échantillons d'individus enfants (0-17 ans) et adultes (18-79 ans) doivent obligatoirement être traités séparément. Il n'est pas possible d'étudier une classe d'âge recoupant les deux échantillons (ex : 15-20 ans).*[4]

Dans notre projet, nous décidons alors de travailler uniquement sur l'échantillon des adultes, c'est-à-dire les individus âgés de 18 à 79 ans.

2.1 Niveaux de participation

Tous les participants n'ont pas suivi le protocole d'étude avec la même intensité. Et compte tenu des différentes étapes et de la multitude des questionnaires posés, l'ANSES a défini trois niveaux de participation pour trois différents groupes:

Groupe	Participation	Effectifs
POP1	● Visite à domicile ● Questionnaire FAF	3157 adultes
POP2	● Visite à domicile ● Complété 2 volets questionnaire AA	2288 adultes
POP3	● Visite à domicile ● Questionnaire AA ● Au moins 2 interviews alimentaires	2121 adultes

Ces niveaux de participation illustrent une autre contrainte de l'étude INCA3:

> *Par ailleurs, selon le volet de questionnaire à étudier, il faut choisir le niveau de participation qui convient.*

En ce qui nous concerne, nous étudierons séparément deux groupes. Car selon leurs implications, les individus n'ont pas répondu aux mêmes questionnaires. Or les données qui nous intéressent sont justement présentes dans des tables différentes, comme nous l'expliquons ci-après.

4https://www.data.gouv.fr/fr/datasets/donnees-de-consommations-et-habitudes-alimentaires-de-letude-inca-3/

Les données INCA3 auxquelles nous avons accès pour notre projet - mises à disposition en open data via le site data.gouv.fr - se présentent sous la forme de 12 tables thématiques accompagnées de notices d'information.

Et comme le permet l"ANSES dans ses recommandations, nous retenons 4 tables pour notre étude, dont la table principale relative à la description des individus:

Tables	Données	Participation
DESCRIPTION_INDIV	Description des individus	Pop1 / Pop2 / Pop3
HABITUDES_INDIV	Habitudes individuelles	Pop2
ACTPHYS_SEDENT	Activité physique et sédentarité	Pop1 / Pop2
CONSO_GPE_INCA3	Consommations	Pop3

Ainsi, pour la population 2 il est possible d'exploiter les tables:

- DESCRIPTION_INDIV
- HABITUDES_INDIV
- ACTPHY_SEDENT

Pour la population 3 il est possible d'exploiter les tables:

- DESCRIPTION_INDIV
- CONSO_GPE_INCA3

2.2 Individus sélectionnés

En résumé, nous étudierons alors dans notre projet les 2 groupes suivants:

- Le groupe **POP2** d'un effectif de **2288 adultes**
- Le groupe **POP3** d'un effectif de **2121 adultes**

Techniquement, chaque individu peut être caractérisé par :

- La tranche d'âge à laquelle il appartient (variable tage_PS)
- Une variable indiquant son niveau de participation (POP2 / POP3)

Par sélection, nous pouvons alors extraire les individus sur lesquels travailler.

3. Sélection des variables

Nos choix de variables sont motivés par les objectifs de notre étude :

- L'observation des données de consommation alimentaire et d'activité physique
- Les notions d'obésité et de surpoids

3.1 Variables sélectionnées pour le groupe POP2

A partir des 3 tables DESCRIPTION_INDIV, HABITUDES_INDIV et ACTPHY_SEDENT :

	Variable	Type	Contribution à l'analyse	Format ou unité	Libellé
1	**NOIND**	QUAL	Technique		Numéro d'individu
2	**pop2 / pop3**	QUAL	Technique		Population
3	**agglo_5cl**	QUAL	Illustrative	1 : Rural 2 : 2 000-19 999 hab. 3 : 20 000-99 999 hab. 4 : >=100 000 hab. 5 : Agglo Paris	Taille d'agglomération
4	**sex_PS**	QUAL	Illustrative	1 Homme 2 Femme	Sexe de l'individu
5	**tage_PS**	QUAL	Illustrative	7 : 18-44 ans 8 : 45-64 ans 9 : 65-79 ans	Age de l'individu
6	**diplome_interv**	QUANT	Active	1: Pas de scolarité - NSP 2: Ecole prim. & Collège 3: CAP-BEP BEPC-brev. 4: Bac. technologique 5: Bac général 6: 1er cycle univ 7: 2ème cycle univ 8: 3ème cycle univ	Individu - Diplôme le plus élevé
7	**revenu**	QUANT	Active	1 <380 €/mois + NSP 2 [380-530[€/mois 3 [530-690[€/mois 4 [690-840[€/mois 5 [840-990[€/mois 6 [990-1 300[€/mois 7 [1 300-1 600[€/mois 8 [1 600-1 900[€/mois 9 [1 900-2 200[€/mois 10 [2 200-2 500[€/mois 11 [2 500-3 100[€/mois 12 [3 100-4 600[€/mois 13 >=4 600 €/mois	Revenu mensuel total du foyer (y c. alloc. sociales, pensions, loyers perçus)
8	**imc**	QUANT	Active	kg/m2	Indice de masse corporelle
9	**fume**	QUAL	Active	1 : Non, mais a déjà fumé 0 : Non, jamais fumé 2 : Oui, occasionnellement 3 : Oui, quotidiennement	Fume

10	**restaurationrapide_freq**	QUANT	Active	fois/jour	Fréq. conso restau rapide
11	**collation_freq**	QUANT	Active	fois/jour	Fréq. collation entre repas
12	**pain_cereales_bio**	QUANT	Active	1: Jamais 2: moins de 1 fois/mois 3: 1 à 3 fois par mois 4: 1 fois par semaine 5: 2 à 4 fois par semaine 6: 5 à 6 fois par semaine 7: Tous les jours	Pain et céréales bio
13	**fruits_legumes_bio**	QUANT	Active	idem	Légumes bio
14	**produits_laitiers_bio**	QUANT	Active	idem	Lait, yaourts, from. bio
15	**viandes_poissons_bio**	QUANT	Active	idem	Viandes, volai., oeufs bio
16	**consommation_bio**	QUAL	Illustrative	0: non consommateur 2: consommateur	Consommateur de produits bio
17	**tv_duree**	QUANT	Active	heure/jour	Durée moy. devant TV
18	**ordi_duree**	QUANT	Active	heure/jour	Durée moy. devant ordi (hors travail)
19	**travail_duree**	QUANT	Active	heure/jour	Durée moy. de travail
20	**sedentarite_duree**	QUANT	Active	heure/jour	Durée moy. activ. sédent
21	**activite_total_duree**	QUANT	Active	heure/jour	Durée totale activ. dom., loisirs et sportives
22	**activite_domloissport_duree**	QUANT	Active	heure/jour	Durée totale activ. dom., loisirs et sportives

3.2 Variables sélectionnées pour le groupe POP3

A partir des 2 tables DESCRIPTION_INDIV et CONSO_GPE_INCA3.

Concernant les variables de la table CONSO_GPE_INCA3, nous choisissons de les regrouper par calcul de moyennes en nous basant sur la classification en pyramide alimentaire du ministère américain de l'Agriculture.[5]

	Variable	Type	Contribution à l'analyse	Format	Libellé
1	**NOIND**	QUAL	Technique		Numéro d'individu
2	**pop2 / pop3**	QUAL	Technique		Population
3	**agglo_5cl**	QUAL	Illustrative	1 : Rural 2 : 2 000-19 999 hab. 3 : 20 000-99 999 hab. 4 : >=100 000 hab. 5 : Agglo Paris	Taille d'agglomération
4	**sex_PS**	QUAL	Illustrative	1 Homme 2 Femme	Sexe de l'individu
5	**tage_PS**	QUAL	Illustrative	7 : 18-44 ans 8 : 45-64 ans 9 : 65-79 ans	Age de l'individu
6	**diplome_interv**	QUANT	Active	1: Pas de scolarité - NSP 2: Ecole prim. & Collège 3: CAP-BEP BEPC-brev. 4: Bac. technologique 5: Bac général 6: 1er cycle univ 7: 2ème cycle univ 8: 3ème cycle univ	Individu - Diplôme le plus élevé
7	**revenu**	QUANT	Active	1 <380 €/mois + NSP 2 [380-530[€/mois 3 [530-690[€/mois 4 [690-840[€/mois 5 [840-990[€/mois 6 [990-1 300[€/mois 7 [1 300-1 600[€/mois 8 [1 600-1 900[€/mois 9 [1 900-2 200[€/mois 10 [2 200-2 500[€/mois 11 [2 500-3 100[€/mois 12 [3 100-4 600[€/mois 13 >=4 600 €/mois	Revenu mensuel total du foyer (y c. alloc. sociales, pensions, loyers perçus)
8	**imc**	QUANT	Active	kg/m2	Indice de masse corporelle
9	**fume**	QUAL	Active	1 : Non, mais a déjà fumé 0 : Non, jamais fumé 2 : Oui, occasionnellement 3 : Oui, quotidiennement	Fume

5https://fr.wikipedia.org/wiki/Pyramide_alimentaire

10	**pain_cereales**	QUANT	Active	g/jour	Pain et céréales
11	**fruits_legumes**	QUANT	Active	g/jour	Fruits et légumes
12	**produits_laitiers**	QUANT	Active	g/jour	Lait, yaourts, fromages
13	**viandes_poissons_oeufs**	QUANT	Active	g/jour	Viandes, volailles, oeufs
14	**produits_sucres**	QUANT	Active	g/jour	Sucres, chocolat, etc.
15	**eau**	QUANT	Active	g/jour	Conso. eaux conditionnées et eau du robinet
16	**alcool**	QUANT	Active	g/jour	Consommations d'alcool

3.3 Variable actives

Nous sélectionnons les variables actives selon des critères de pertinence en relation avec le thème étudié, ainsi que leur représentation dans l'ensemble du jeu de données. Ces variables - quantitatives continues ou discrètes - sont normalisées avant de participer à la construction des axes de notre ACP.

- **Groupe POP2** :
  revenu, imc, restaurationrapide_freq, collation_freq, fruits_legumes_bio, tv_duree, travail_duree, activite_total_duree
- **Groupe POP3** :
  diplome_interv, revenu, imc, pain_cereales, fruits_legumes, produits_laitiers, viandes_poissons_oeufs, produits_sucres, eau, alcool

3.4 Variables illustratives

Nous sélectionnons certaines variables illustratives - qualitatives ou discrètes à peu de modalités - comme aide possible à l'interprétation des résultats de l'ACP. Ces variables ne participent pas à la construction des axes.

- **Groupe POP2** : *sex_PS, tage_PS, diplome_interv, fume, consommation_bio*
- **Groupe POP3** : *agglo_5cl, sex_PS, tage_PS, diplome_interv, fume*

4. Pré- traitement des données

4.1 Importation et traitement des tables

Tables retenues pour la population 2 (**2288 adultes):**

- DESCRIPTION_INDIV
- HABITUDES_INDIV
- ACTPHY_SEDENT

Tables retenues pour la population 3: (**2121 adultes**)

- DESCRIPTION_INDIV
- CONSO_GPE_INCA3

4.2 Gestion des valeurs manquantes

Divers traitements de sélection des variables et individus sont effectués sur les tables importées. Deux dataframes généraux pour chaque groupe POP2 et POP3 sont constitués par jointures.

Les données sont vérifiées par affichages et contrôles grâce à une fonction Python/Pandas *describe()* appliquée aux variables quantitatives. Les valeurs manquantes sont remplacées par les moyennes pour ces variables.

Les dataframes résultats des notebooks sont sauvegardés dans deux fichiers distincts :

- pop2_df_base.csv
- pop3_df_base.csv

4.3 Normalisation des données

Ces fichiers df_base.csv sont eux-mêmes importés dans deux autres notebooks distincts dans lesquels seront effectués les prétraitements et calculs d'ACP et classifications.

- pop2_acp.ipynb
- pop3_acp.ipynb

La normalisation des données - centrage et réduction - est effectuée grâce à la classe *StandardScaler* de la librairie Python Sklearn. Le centrage étant une soustraction des moyennes aux variables, et la réduction étant une division par écart-type. Cette mise à l'échelle va nous permettre d'effectuer un calcul d'ACP sur ces données.

B. Analyse statistique

1. Analyse univariée

1.1 Groupe POP2 - Informations de base : moyennes, écart-types et quartiles, etc.

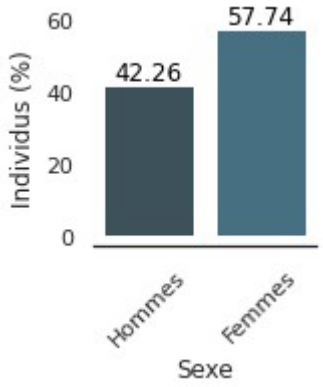
	imc	tv_duree	ordi_duree	travail_duree	sedentarite_duree
count	2288.000000	2288.000000	2288.000000	2288.000000	2288.000000
mean	25.793700	3.213525	1.870503	3.506600	6.349799
std	4.930130	1.760040	1.421218	1.711142	3.216838
min	15.776830	0.142860	0.077308	0.035714	0.142860
25%	22.320882	2.000000	1.000000	2.429770	3.999996
50%	25.037310	3.000000	1.585665	2.429770	6.000000
75%	28.394135	4.000000	2.250825	5.000000	8.464276
max	52.608677	9.000000	9.000000	14.271429	16.997459

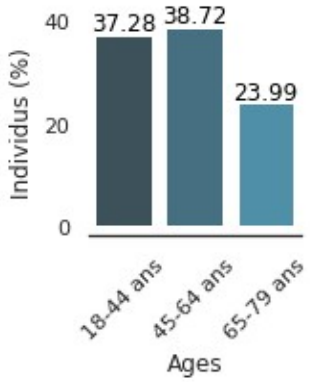
	activite_total_duree	activite_domloissport_duree
count	2288.000000	2288.000000
mean	10.123324	2.398787
std	3.607025	2.146942
min	0.119044	0.002976
25%	7.644349	1.057664
50%	10.118902	1.773805
75%	2.416915	3.053947
max	17.000002	25.830357

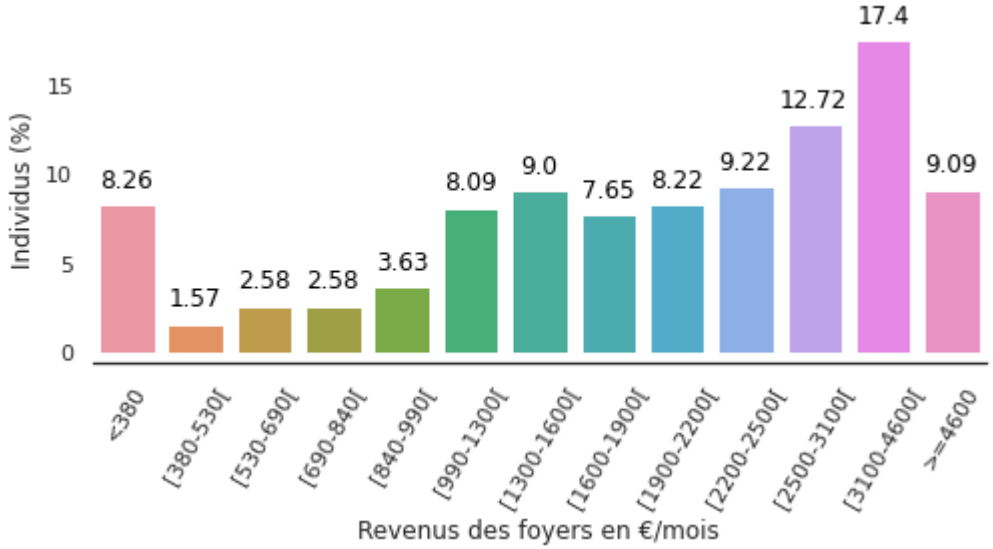
	pain_cereales_bio	fruits_legumes_bio	produits_laitiers_bio	viandes_poissons_bio
count	2288.000000	2288.000000	2288.000000	2288.000000
mean	0.350962	0.392045	0.695367	0.496066
std	0.677961	0.707674	0.928269	0.736328
min	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000
50%	0.000000	0.000000	0.000000	0.000000
75%	1.000000	1.000000	1.000000	1.000000
max	3.000000	3.000000	3.000000	3.000000

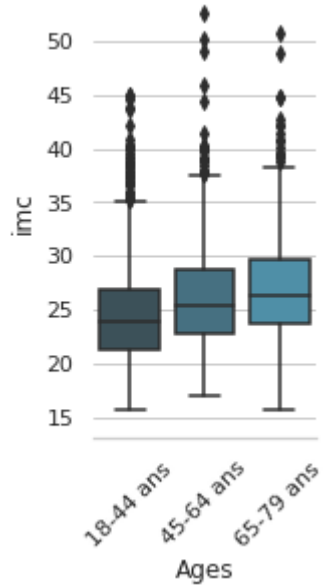
	restaurationrapide_freq	collation_freq
count	2288.000000	2288.000000
mean	1.569073	2.624239
std	0.809681	0.975023
min	1.000000	1.000000
25%	1.000000	2.000000
50%	1.184441	2.066871
75%	2.000000	3.000000
max	6.000000	5.000000

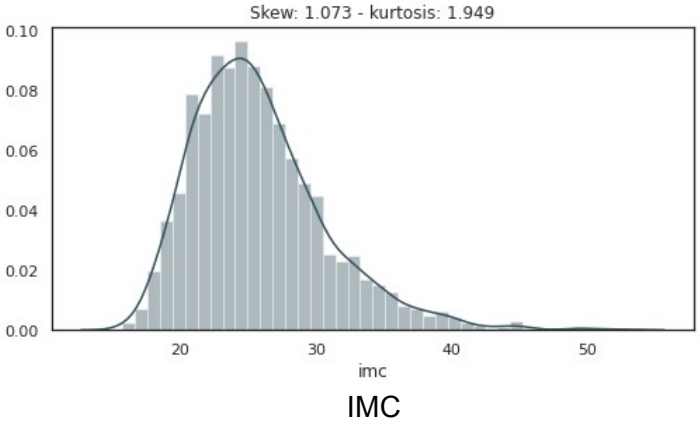
1.2 Groupe POP2 - Informations de base : Sexe, Âges, Revenus, IMC, Sédentarité











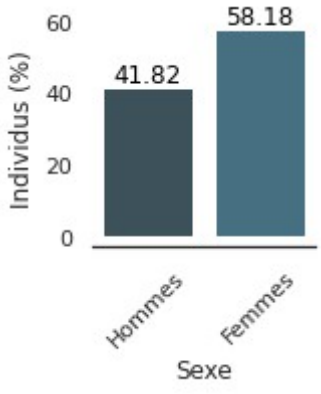
IMC

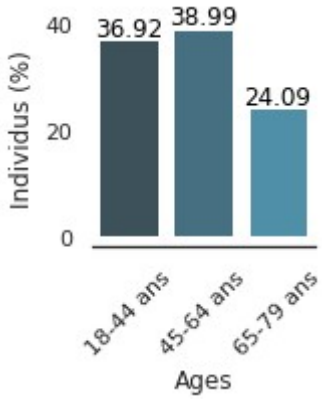
1.3 Groupe POP3 : moyennes, écart-types et quartiles, etc.

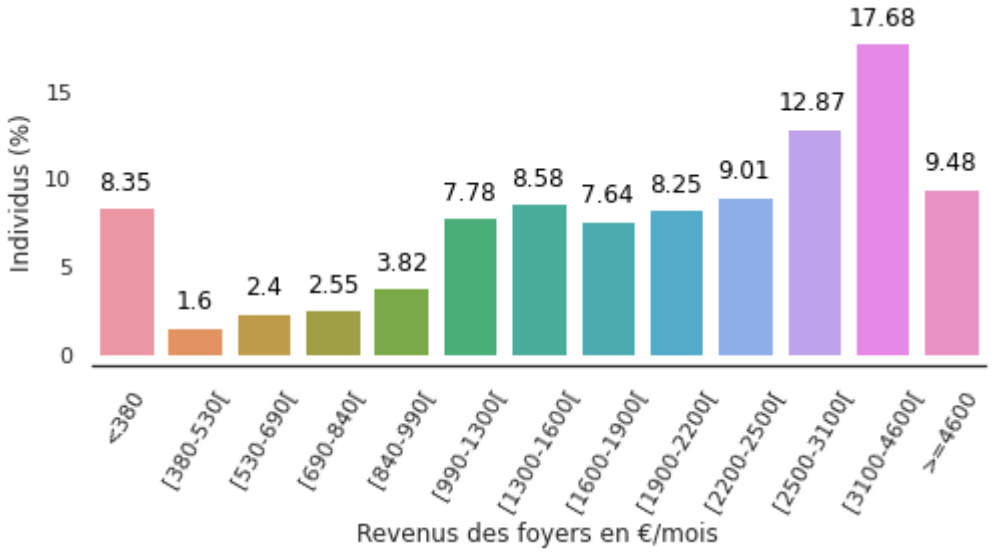
	imc	pain_cereales	fruits_legumes	viandes_poissons_oeufs
count	2121.000000	2121.000000	2121.000000	2121.000000
mean	25.768907	114.817111	157.652167	57.376445
std	4.812153	88.125929	116.004836	48.907099
min	15.776830	0.535714	1.010000	0.128143
25%	22.321428	52.540625	74.000000	23.107144
50%	25.050505	96.244995	140.375007	51.785294
75%	28.405504	152.250000	211.428570	73.528908
max	50.117188	644.872680	962.928590	400.714290

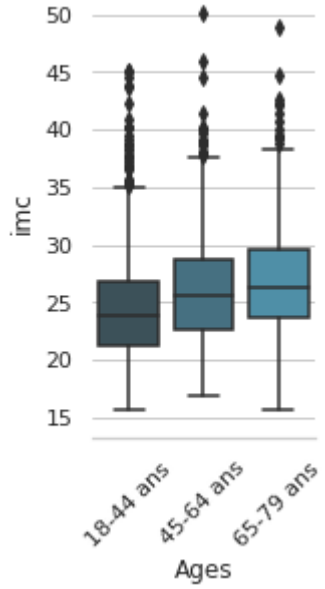
	produits_laitiers	produits_sucres	eau	alcool
count	2121.000000	2121.000000	2121.000000	2121.000000
mean	177.531658	84.080052	912.271482	184.590626
std	144.294283	63.100138	579.819760	226.407082
min	1.442857	1.428572	6.696428	0.735714
25%	89.285713	41.321426	482.991090	102.973230
50%	149.958043	72.024956	815.714330	133.078422
75%	209.748810	108.571423	1222.499996	175.714280
max	2007.288030	447.674010	3896.473400	6209.856900

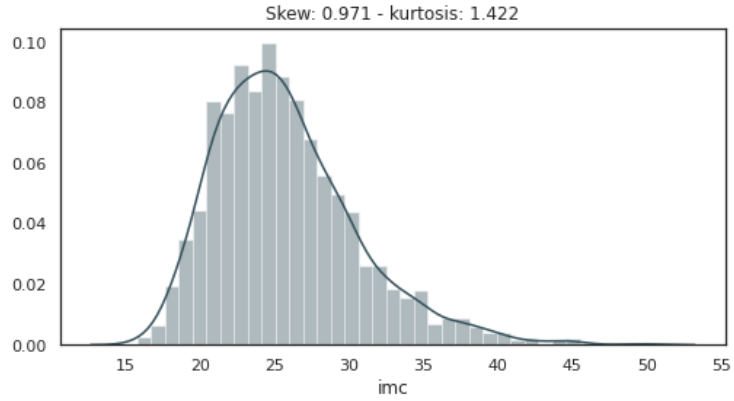
1.4 Groupe POP3 : Sexe, Âges, Revenus, IMC, Sédentarité











1.5 Analyse univariée - Interprétations

Les deux groupes étudiés (POP2 et POP3) présentent des répartitions d'individus relativement similaires avec une majorité de femmes (16% de plus que d'hommes). Les classes d'âges sont homogènes; la classe d'âge la plus élevé étant la moins nombreuse.

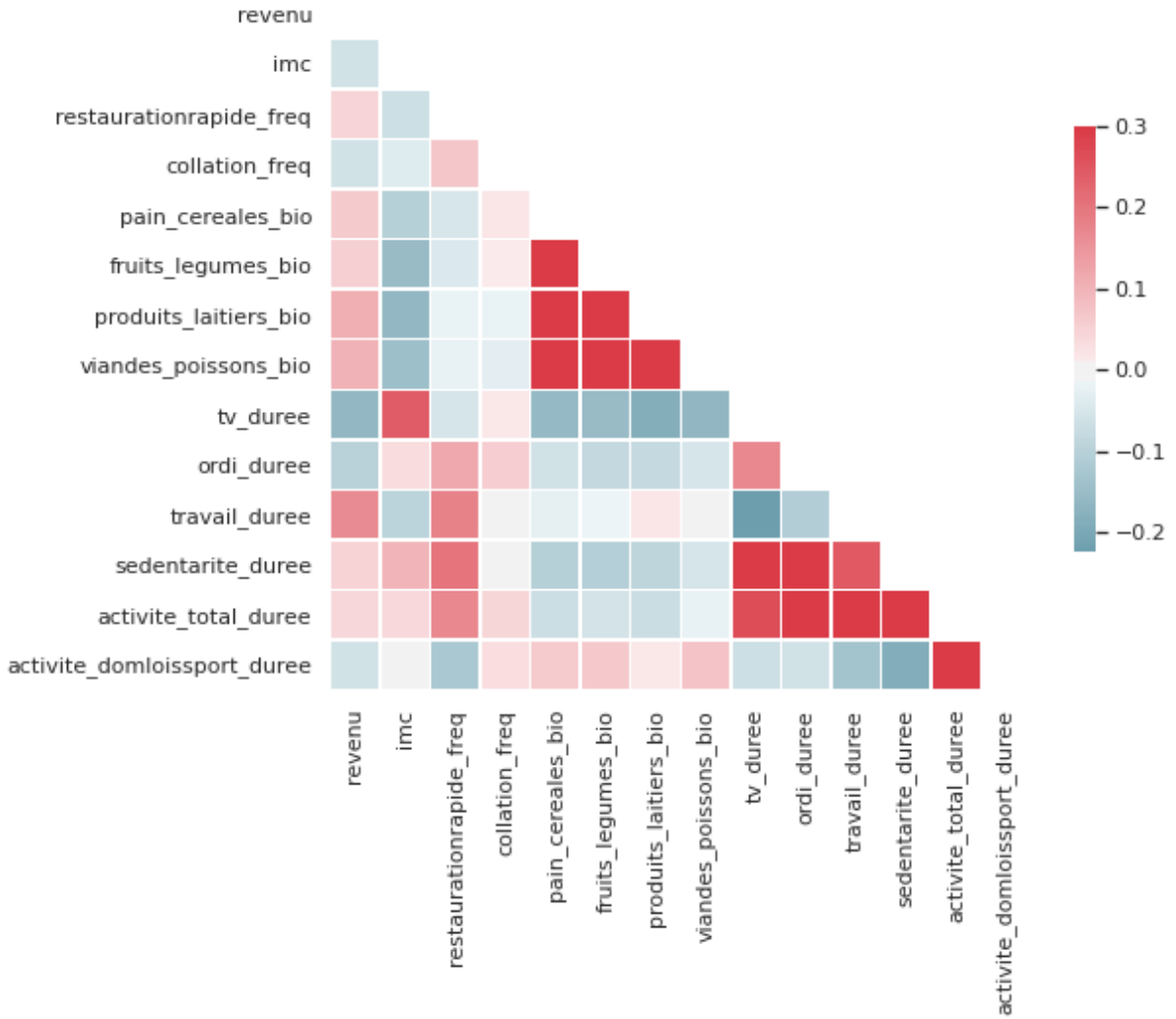
Les distributions de revenus sont asymétriques. Les calculs de coefficients d'asymétrie de Fisher (-0.705 pour POP2 et -0.720 pour POP3) indiquent des queues de distributions à gauche. Les coefficients d'aplatissement (-0.514 pour POP2 et -0.504 pour POP3) signifient des distributions plus aplaties que la Gaussienne. Sont à noter - visuellement - trois pics dans ces diagrammes de revenus, qui semblent définir trois groupes d'individus.

Dans les deux groupes POP2 et POP3, l'indice de masse corporelle - imc - présente une distribution plus concentrée ou pointue que la Gaussienne, avec des coefficients d'aplatissement de Fisher de 1.073 pour POP2 et 0.971 pour POP3. Les coefficients d'asymétrie de Fisher - de 1.949 et 1.422 indiquent des queues de distributions à droite.

2. Analyse bivariée

2.1 Matrice des corrélations - groupe POP2

A l'aide de la fonction *corr(method='pearson')* des librairies Pandas et Seaborn, nous construisons la matrice des corrélations des variables quantitatives du groupe POP2.



Se distinguent deux groupes de variables corrélées entre elles (coefficients proches de +0.3): Un groupe autour des consommations de produits bio, et un autre relatif à la sédentarité ou activité physique. Nous comprenons par exemple que les variables tv_duree et sedentarite_duree augmentent ensembles.

La liaison entre imc et tv_duree est également notable. Il semblerait aussi que les variables tv_duree et revenu évoluent en sens contraires. Observons par exemple les résultats de corrélations entre la variables imc revenu, restaurationrapide_freq et les autres :

	imc	revenu	restaurationrapide_freq
revenu	-0.058637	1.000000	0.047892
imc	1.000000	-0.058637	-0.065057
restaurationrapide_freq	-0.065057	0.047892	1.000000
collation_freq	-0.034055	-0.059593	0.072464
pain_cereales_bio	-0.101743	0.063102	-0.049156
fruits_legumes_bio	-0.152221	0.058012	-0.043387
produits_laitiers_bio	-0.160485	0.108378	-0.020814
viandes_poissons_bio	-0.143287	0.104765	-0.022519
tv_duree	**0.241430**	-0.160544	-0.053279
ordi_duree	0.036450	-0.099363	0.118202
travail_duree	-0.095322	**0.168463**	0.180350
sedentarite_duree	0.099624	0.050035	**0.205163**
activite_total_duree	0.041518	0.043568	0.170937
activite_domloissport_duree	0.005860	-0.060457	-0.123610

- La variable avec laquelle l'imc est le plus corrélée est tv_duree.
- Pour la variable revenu le maximum de corrélation se trouve avec travail_duree.
- restaurationrapide_freq est corrélée positivement avec sedentarite_duree.

A ce stade de notre étude, ces informations nous indiquent simplement qu'une augmentation de la variable tv_duree entraine une élévation de l'imc. Le fait que le revenu augmente avec travail_duree - action prévisible et heureuse - ne nous intéresse pas ici (à ce niveau d'édude).

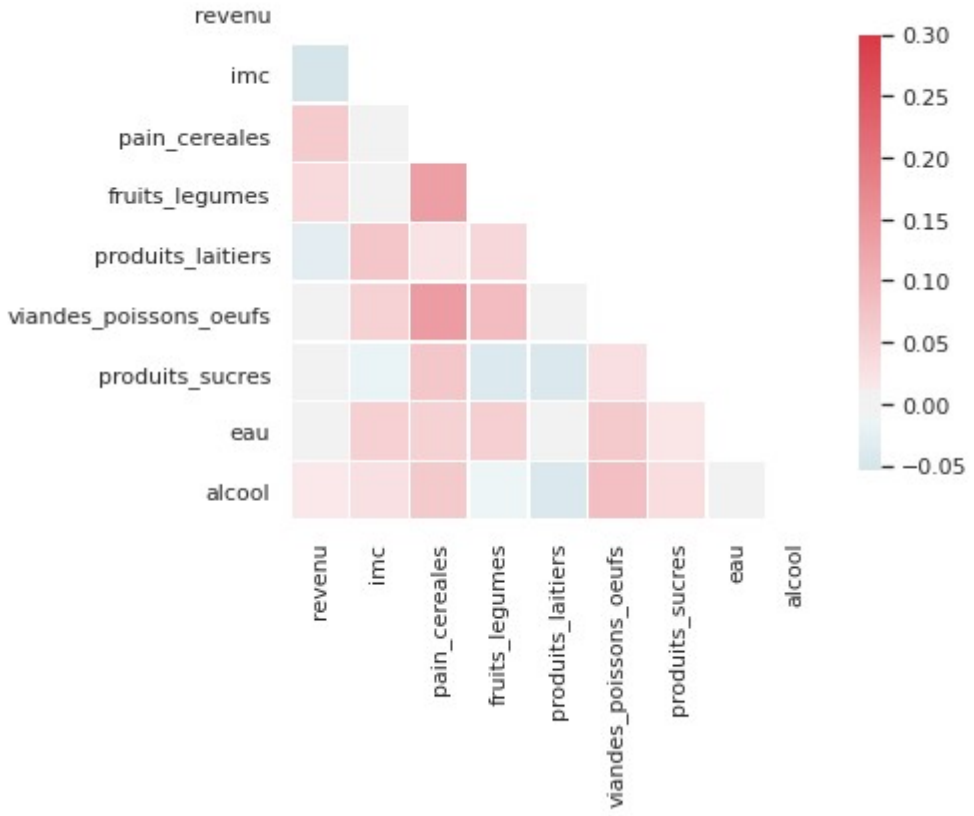
Cependant la corrélation entre les variables restaurationrapide_freq et sedentarite_duree nous semble intéressante à noter. Pour la caractéristique de la variable restaurationrapide_freq d'une part, et pour sa liaison avec sedentarite_duree.

En décision de cette étape d'analyse bivariée pour POP2, nous décidons de ne retenir - au niveau de l'ACP - que certaines variables représentatives des groupes:

- Nous choisirons *fruits_legumes_bio* pour les variables de produits bio
- Nous choisirons *tv_duree* pour les variables de sédentarité

2.2 Matrice des corrélations - groupe POP3

De manière similaire au paragraphe précédent, nous construisons - à l'aide de la fonction *corr(method='pearson')* des librairies Pandas et Seaborn - la matrice des corrélations des variables quantitatives du groupe POP3.



Les corrélations les plus fortes apparaissent entre les couples de variables *fruits_legumes* / *pain_cereales* et *viandes_poissons_oeufs* / *pain_cereales*. En notant également une corrélation entre les consommations de viandes, poissons, oeufs et fruits et légumes, nous constatons que ces quantités de consommations évoluent dans le même sens.

Les valeurs minimum de coefficients (-0.05) sont également notables. Les variables évoluant en sens contraires le font de manière plus modérée. A remarquer - par exemple - la corrélation négative entre IMC et revenu qui nous fait penser que plus les revenus augmentent, plus les IMC des individus diminuent.

Nous retenons également la corrélation négative entre *produits_sucres* et fruits_legumes ou *produits_laitiers*. De même sont à retenir les faibles corrélations des couples de variables *eau* / *revenu* et *alcool* / *fruits_legumes*.

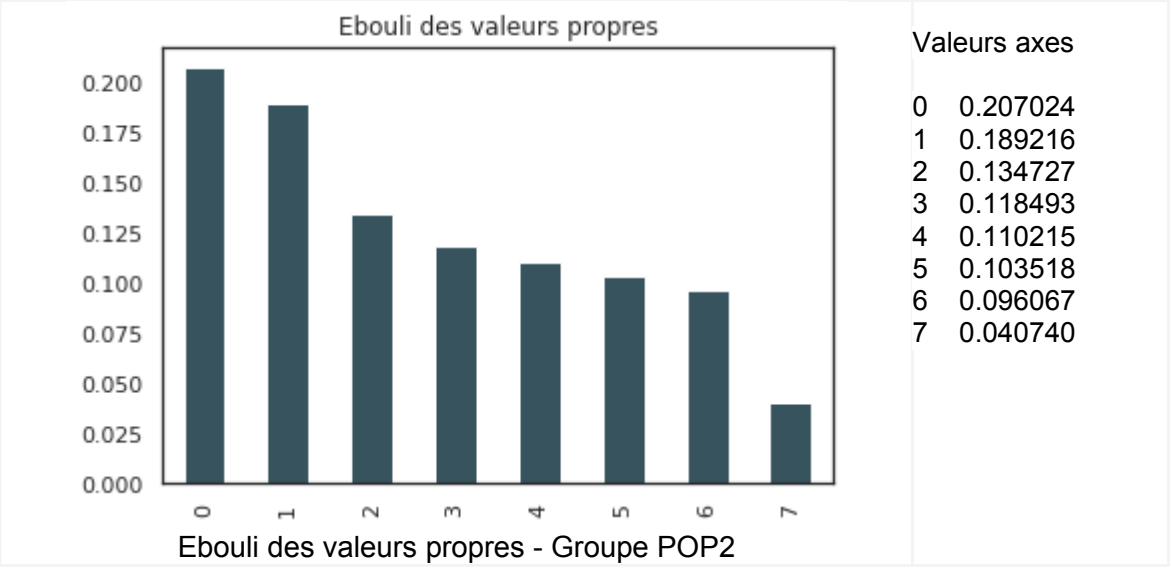
3. Analyses en composantes principales

Nous cherchons à synthétiser l'information de notre jeu de données en regroupant les variables liées en de nouvelles variables synthétiques. L'ACP va nous permettre d'étudier la variabilité des individus, c'est-à-dire leurs ressemblances et différences ainsi que les liaisons entre variables.

Les données - préalablement traitées pour les cas de valeurs manquantes - sont centrées et réduites grâce à la classe *StandardScaler* de la librairie scikit-learn. Nous utiliserons ensuite la classe *PCA* de cette même librairie Python.

Nous projetons nos données sur les axes principaux d'inertie qui sont ordonnés selon l'inertie du nuage projeté. Nous décrivons le pourcentage d'inertie totale associée à chaque axe à l'aide du diagramme "ébouli des valeurs propres" appliqué sur chacun de nos jeux de données. Il consiste en une représentation en pourcentage de variance expliquée, grâce à la fonction *explained_variance_ratio_* de la classe PCA de la librairie scikit-learn.

3.1. Choix du nombre d'axes - Groupe POP2



Valeurs axes

0	0.207024
1	0.189216
2	0.134727
3	0.118493
4	0.110215
5	0.103518
6	0.096067
7	0.040740

Ebouli des valeurs propres - Groupe POP2

Nous cherchons à savoir ici combien de composantes principales analyser: L'apport des deux premiers axes est le plus important en termes d'explication de la variance, et la "règle du coude" - cassure dans l'ébouli des valeurs propres - nous entraînerait à les retenir exclusivement, mais ils ne représentent que 28% de l'inertie totale.

Nous décidons alors d'analyser 6 composantes qui représentent 75% de l'inertie totale.

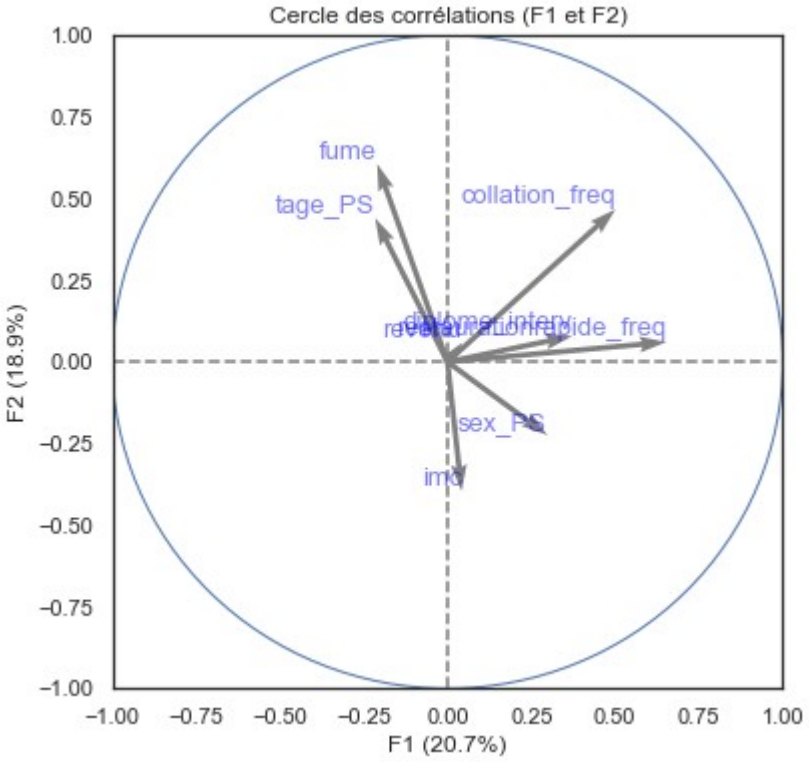
3.2. Cercles des corrélations - Groupe POP2

La projection du nuage des variables sur le premier plan factoriel nous permet l'interprétation des axes et des variables contributrices pour le groupe POP2.

POP2

Parmi les variables les mieux représentées dans ce premier plan factoriel, *collation_freq* et *restaurationrapide* sont corrélées à l'axe **F1**. Leur notion commune pourrait être :
MALBOUFFE

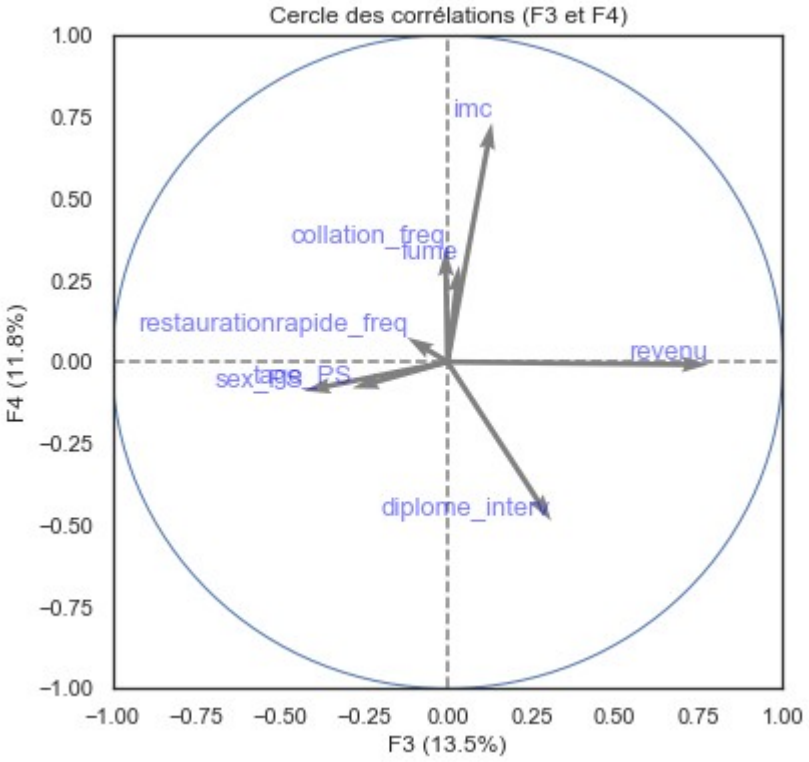
Les variables *fume* - relativement bien représentée - est corrélée à l'axe **F2** que nous pouvons définir par:
FUMEUR



POP2

La variable *revenu* est bien projetée sur **F3**, donc fortement corrélée à cet axe. Nous retenons la notion de :
RESSOURCE.

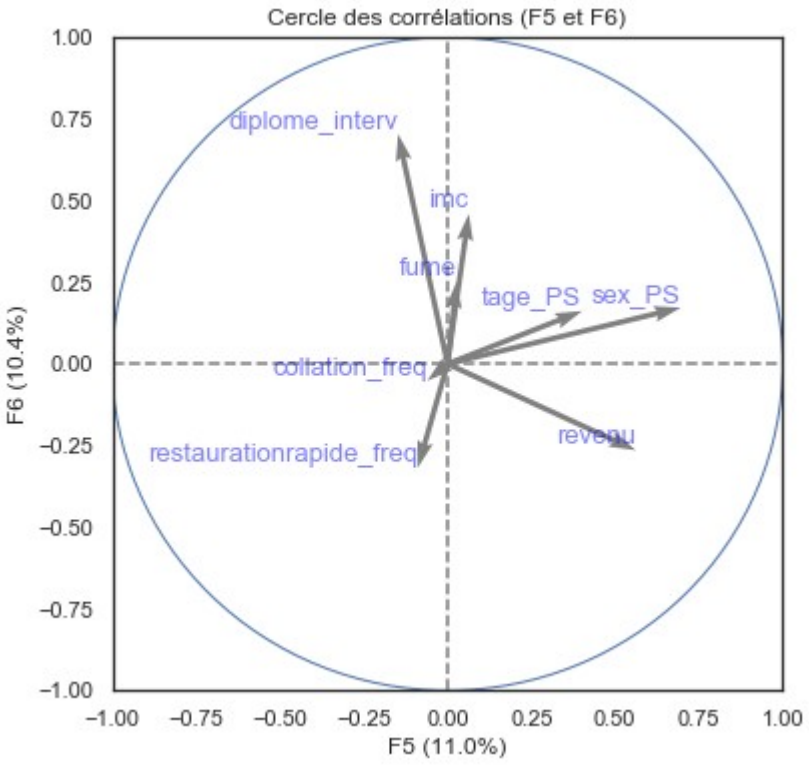
La variable *imc* est bien projetée et corrélée à l'axe **F4** que nous pouvons identifier par la notion de:
SANTE



POP2

Les variables *sex_PS* et *tage_PS* sont corrélée à l'axe **F5**, qui peut être décrit par la notion de: ***SOCIO-DEMO***

Enfin la variable diplome_interv est bien projetée et donc corrélée sur l'axe **F6**. Nous retenons la notion de : ***EDUCATION***



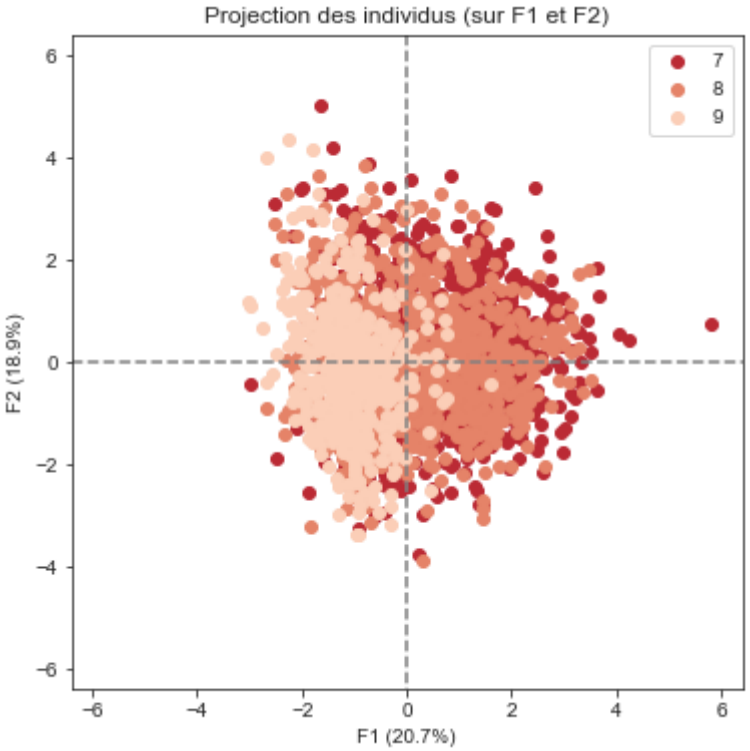
3.3. Représentations des individus sur les plans factoriels - Groupe POP2

Avec ces graphes - colorés en fonction des revenus - nous souhaitons analyser les différences et similarités entre individus, ainsi que leurs positions relatives aux axes.

Graphe des individus POP2
F1→ ***MALBOUFFE***
F2→ ***FUMEUR***

Le nuage de points semble dessiner un triangle incliné dont à la pointe seraient placés les fumeurs qui s'alimentent sainement et à la base les non-fumeurs adeptes de malbouffe.

Variable âge (indicatif)
7 : 18-44 ans
8 : 45-64 ans
9 : 65-79 ans

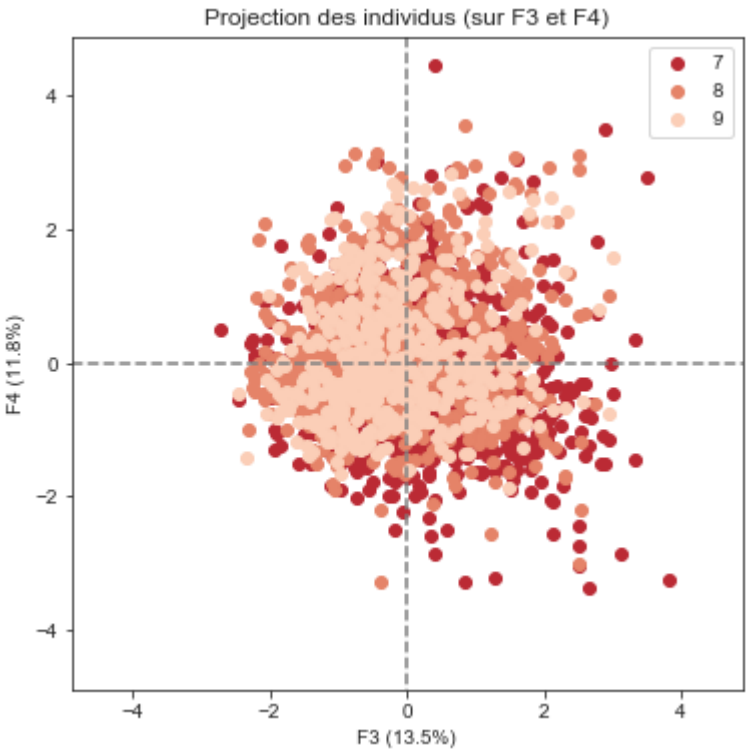


[Les individus sont colorés selon tranches d'âges.]

Graphe des individus POP2
F3 → ***RESSOURCE***
F4 → ***SANTE***

Nous observons une relative symétrie dans le nuage de points de part et d'autre de l'axe F4 /SANTE.

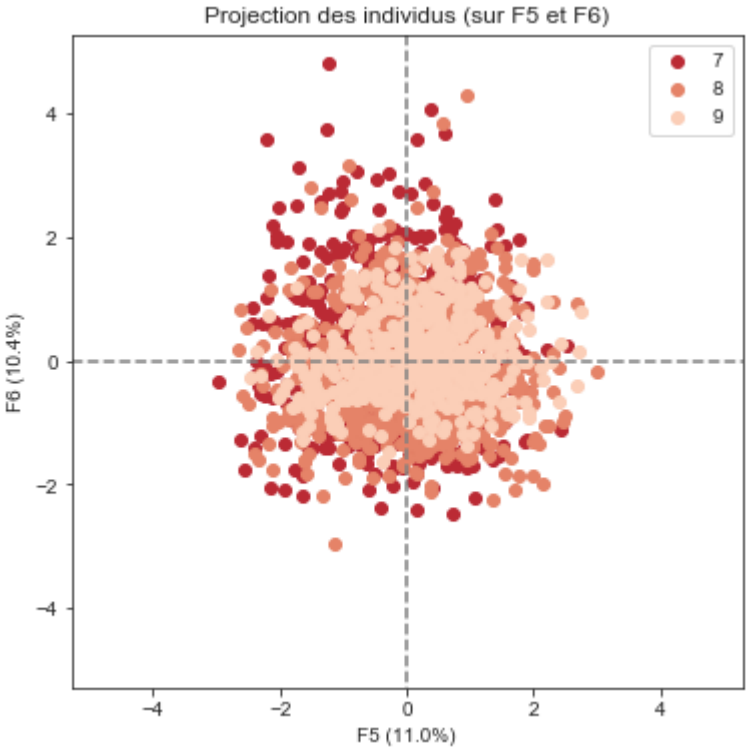
Variable âge (indicatif)
7 : 18-44 ans
8 : 45-64 ans
9 : 65-79 ans



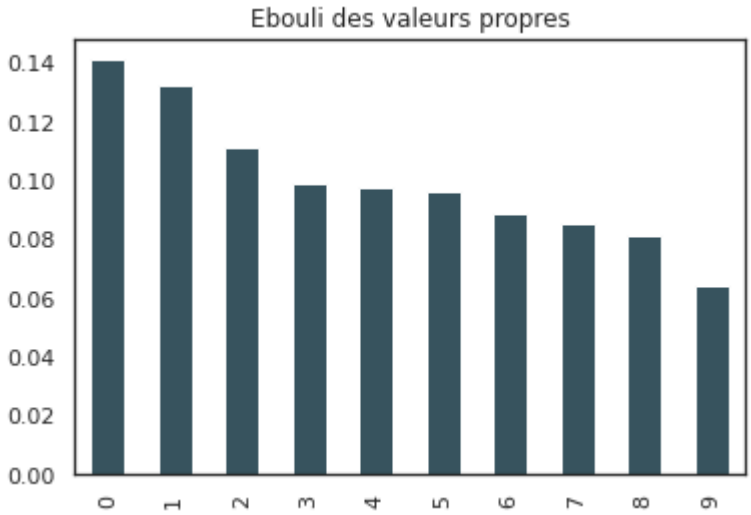
Graphe des individus POP2
F5 → ***SOCIO-DEMO***
F6 → ***EDUCATION***

Sur cette projection, le nuage de point est relativement symétrique autour des 2 axes, avec davantage d'outliers vers le haut, axe F6 / EDUCATION.

Variable âge (indicatif)
7 : 18-44 ans
8 : 45-64 ans
9 : 65-79 ans



3.4. ACP - Groupe POP3 - Choix du nombre d'axes



Ebouli des valeurs propres - Groupe POP3

Valeurs axes

0	0.141317
1	0.132602
2	0.111316
3	0.099481
4	0.097890
5	0.096234
6	0.089255
7	0.085768
8	0.081702
9	0.064435

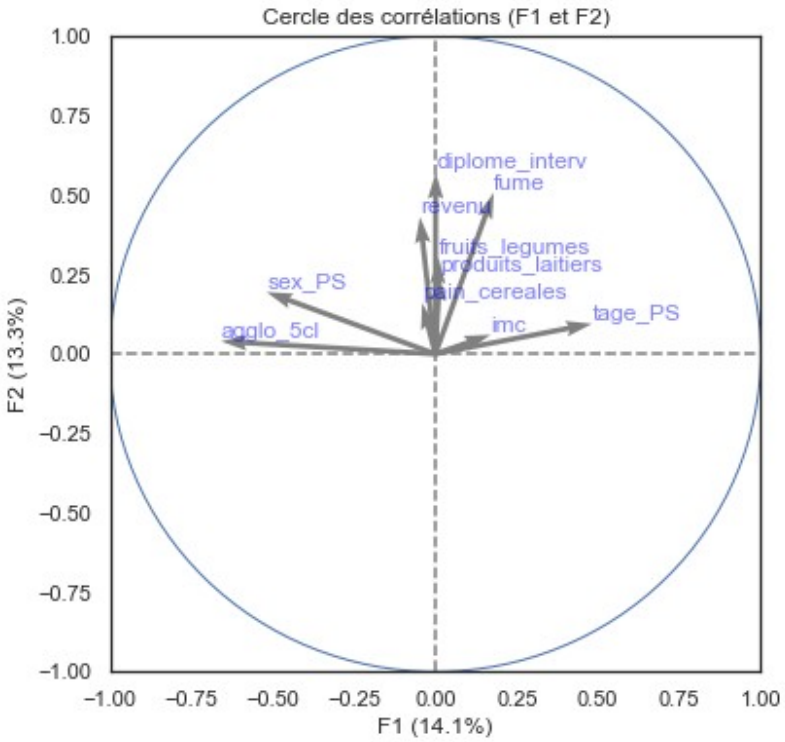
Dans le cas du groupe POP3, l'apport des deux premiers axes - bien que le plus important - ne représente que 27% de l'inertie totale. Nous décidons alors d'analyser 6 composantes qui représentent **68%** de l'inertie totale.

3.5. Interprétation des axes et des variables contributrices - Groupe POP3

POP3

Les variables *tage_PS*, *agglo_5cl* et *sex_PS* sont corrélées à l'axe **F1** que nous pouvons décrire par la notion de : ***SOCIO-DEMO***

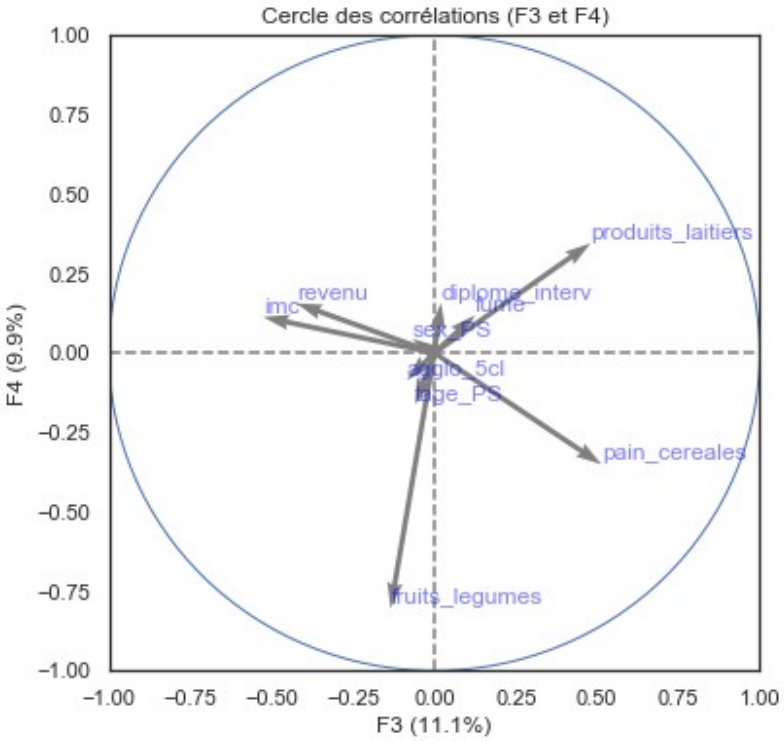
Les variables *diplome, revenu* et *fume* sont corrélées à **F2**. Malgré l'absence de notion commune nous retenons pour cet axe l'idée de : ***EDUCATION***



POP3

Les variables *produits_laitiers* et *pain_cereales*, sont corrélés à F3 que nous décrivons par la notion de :
PAIN-CEREALES

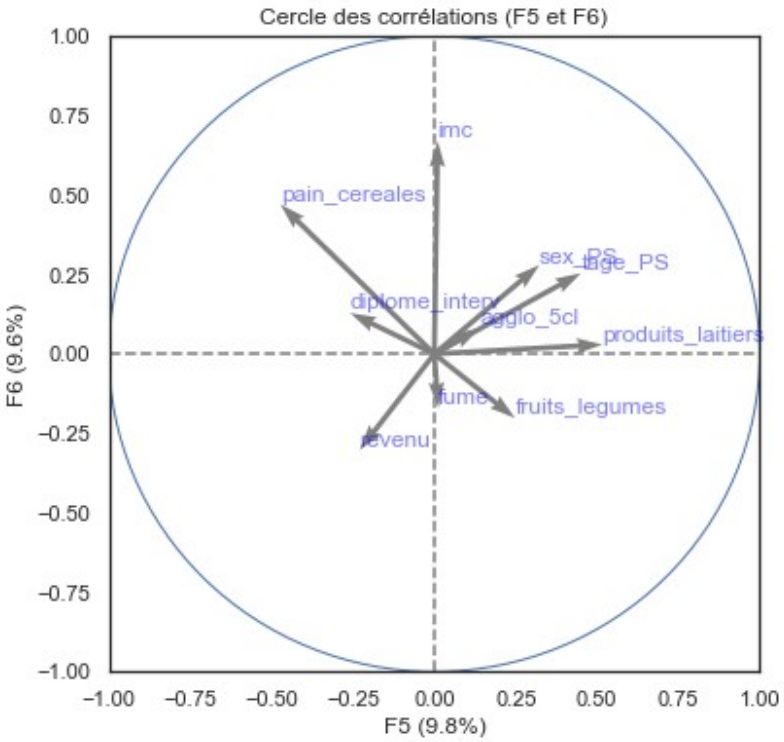
La variable *fruits_legumes* est fortement corrélée à l'axe **F4** que nous désignons par la notion:
FRUITS-LEGUMES



POP3

La variable *produits_laitiers* est bien projetée sur l'axe **F5** que nous désignons par la notion de :
PRODUITS-LAITIERS

La variable *imc* est corrélée à l'axe **F6** pour lequel nous retenons la notion de :
SANTE



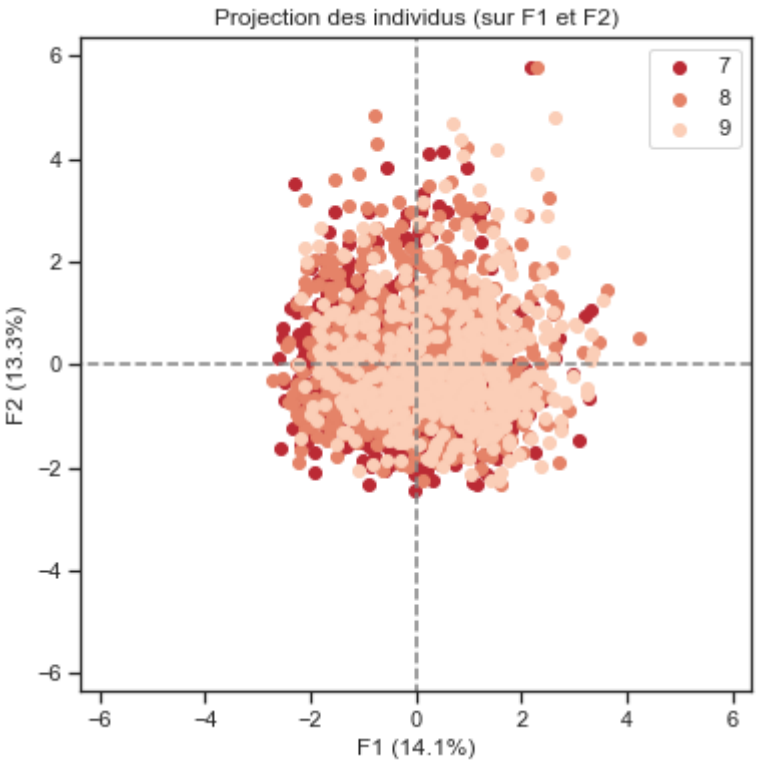
3.6. Représentations des individus sur les plans factoriels - Groupe POP3

Graphe des individus POP3
F1 → ***SOCIO-DEMO***
F2 → ***EDUCATION***

Le nuage des individu montre une symétrie circulaire avec un certain étirement vers le haut, selon la notion EDUCATION / axe F1.

Les différentes classes sociales sont visibles en relation à l'axe des abscisses, désigné par la notion socio-démographique.

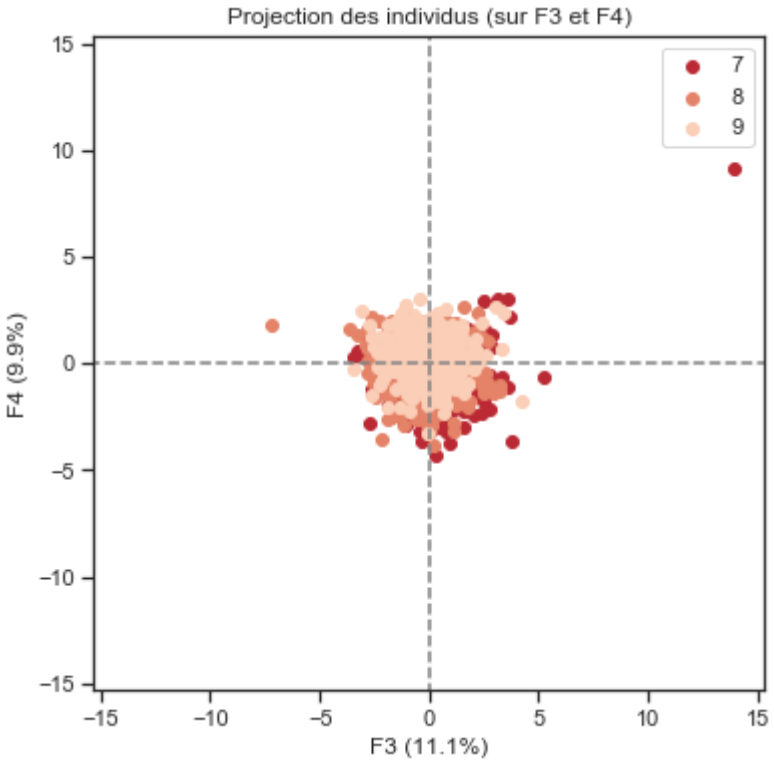
[Les individus (points) sont colorés selon niveaux de revenus.]



Graphe des individus POP3
F3 → **PAIN-CEREALES**
F4 → **FRUITS-LEGUMES**

Les individus sont ici agglomérés de manière très homogène et symétrique autour des notions pain-céréales et fruits-légumes.

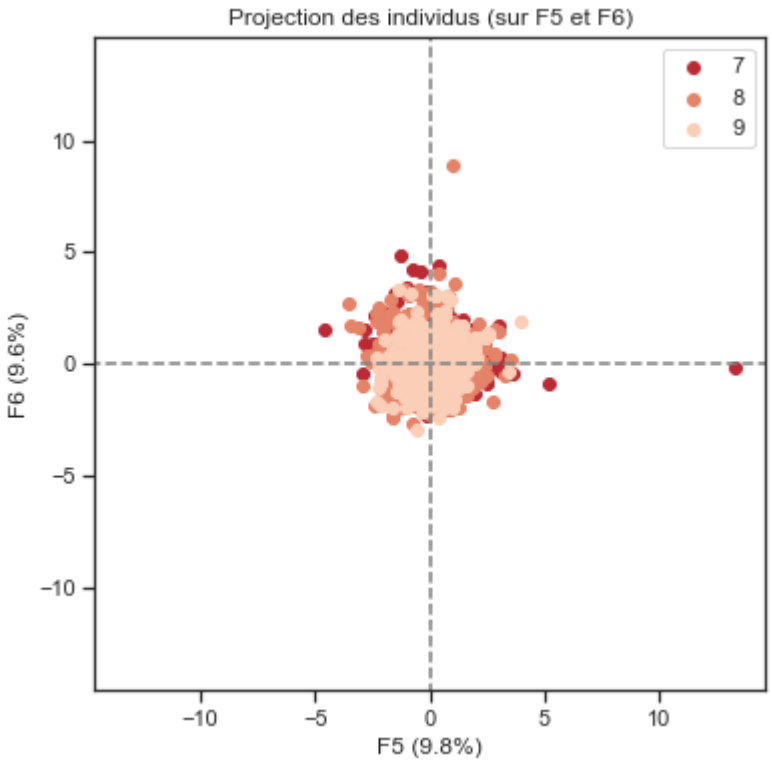
Variable âge (indicatif)
7 : 18-44 ans
8 : 45-64 ans
9 : 65-79 ans



Graphe des individus POP3
F5 → **PRODUITS-LAITIERS**
F6 → ***SANTE***

On observe également une agglomération des individus autour des notions de produits laitiers et santé.

Variable âge (indicatif)
7 : 18-44 ans
8 : 45-64 ans
9 : 65-79 ans



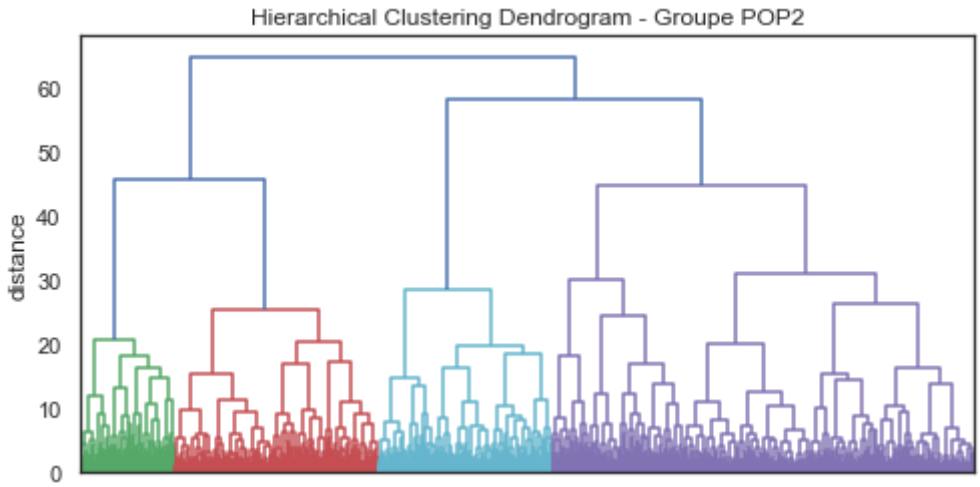
4. Classifications Ascendante Hiérarchique et K-means

Nous souhaitons maintenant étudier de possibles groupes ou clusters parmi nos individus à l'aide de méthodes de classification. La CAH - classification ascendante hiérarchique - nous donnera une idée sur le nombre de clusters à définir. Elle produit une structure - arborescence - permettant la mise en évidence de liens hiérarchiques entre individus ou groupes d'individus. La CAH permet également la détection d'un nombre de classes "naturel" au sein de la population. En effet son algorithme ne requiert pas de nombre de classes au démarrage, contrairement à l'algorithme k-means.

Nous produisons une CAH sur notre jeu de données grâce à la fonction *linkage* de la librairie scikit-learn - *linkage(X_scaled, 'ward')* -. En argument de fonction nous spécifions l'emploi de la méthode de Ward qui, à chaque itération - c'est-à-dire à chaque fois que 2 clusters sont regroupés en 1 - cherche à minimiser l'augmentation d'inertie intraclasse due au regroupement des 2 clusters.

Nous consoliderons ensuite nos partitionnements avec l'algorithme k-means qui nous permettra l'interprétation des groupes d'individus. La méthode k-means - ou méthode d'agrégation autour des centres mobiles - ne nécessite pas la construction d'un arbre hiérarchique. L'inconvénient étant la perte des liens hiérarchiques entre individus.

4.1. Partitionnement par CAH et k-means - Groupe POP2



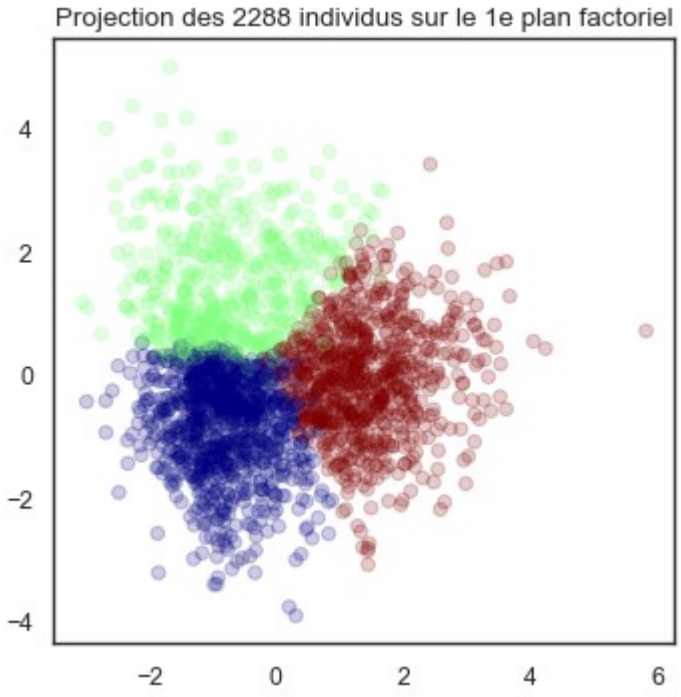
Sur le groupe POP2, le dendrogramme issu d'une CAH, nous suggère un partitionnement en 3 groupes. En effet, nous souhaitons récupérer le plus possible de variabilité totale c'est-à-dire conserver un maximum d'information. Un nombre trop faible de classe nous conduirait à des classes qui ne sont pas homogènes. Tandis que construire une partition avec trop de classes risquerait de conduire à des classes qui ne se différencient pas suffisamment.

Notre choix de 3 groupes distincts repose sur la volonté de couper l'arbre hiérarchique dans une partie où les branches sont assez longues. Enfin, notre dernier critère se base sur l'interprétabilité des classes, que nous allons traiter avec l'algorithme k-means.

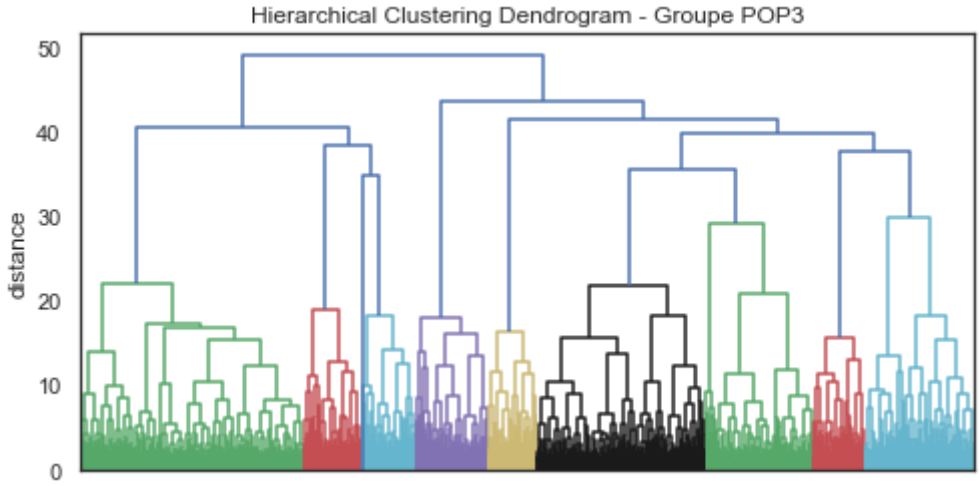
Le partitionnement avec l'algorithme k-means - classe *KMeans* de scikit-learn - illustre 3 groupes d'individus clairement homogènes.

Rappelant les notions:
F1→ ***MALBOUFFE***
F2→ ***FUMEUR***

nous suggérons ces groupes:
- **Pro-malbouffe** (à droite)
- **Anti-malbouffe** (à gauche)
- **Fumeurs** (en haut)



4.2 Partitionnement par CAH et k-means - Groupe POP3

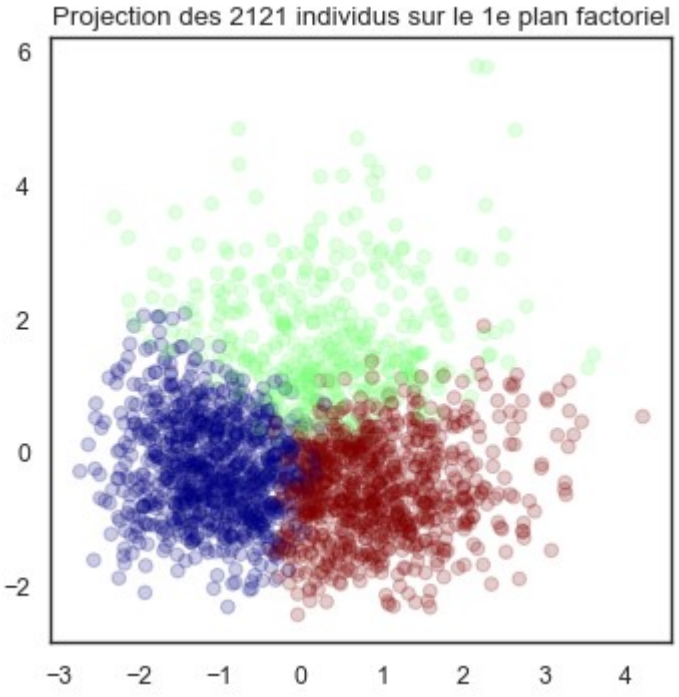


Sur le groupe POP3, le dendrogramme issu d'une CAH présente des branches hautes peu longues et suggère des groupes homogènes plus nombreux que dans le cas de POP2. Dans un soucis de conserver un maximum d'information, obtenir des groupes homogènes et pouvoir les interpréter, nous choisissons de retenir cependant un nombre de 3 clusters et de les observer avec le k-means.

Le partitionnement en 3 classes à l'aide de l'algorithme k-means illustre 3 groupes d'individus clairement homogènes.

Rappelant les notions:
F1 → ***SOCIO-DEMO***
F2 → ***EDUCATION***

nous suggérons ces groupes:
- **Plus âgés** (à droite)
- **Moins âgés** (à gauche)
- **Informés** (en haut)



Conclusion

Ce projet nous a permis d'explorer et synthétiser les informations issues de deux groupes d'individus en résumant les variables à 6 notions distinctes:

POP2	POP3
F1 → ***MALBOUFFE***	F1 → ***SOCIO-DEMO***
F2 → ***FUMEUR***	F2 → ***EDUCATION***
F3 → ***RESSOURCE***	F3 → ***PAIN-CEREALES***
F4 → ***SANTE***	F4 → ***FRUITS-LEGUMES***
F5 → ***SOCIO-DEMO***	F5 → ***PRODUITS-LAITIERS***
F6 → ***EDUCATION***	F6 → ***SANTE***

Nous discernons également 3 clusters principaux d'individus parmi nos jeux de données:

POP2	POP3
● **Pro-malbouffe**	● **Plus âgés**
● **Anti-malbouffe**	● **Moins âgés**
● **Fumeurs**	● **Informés**

Notre point de départ était l'exploration des causes du surpoids et de l'obésité au sein d'une tranche de la population française. L'ACP nous a permis de découvrir les différentes facettes des liens entre les variables. Aussi, malgré les difficultés rencontrées comme les faibles qualités de représentations de certaines variables, nous pouvons noter l'importance de certaines notions telles que la "malbouffe", le fait de fumer ou non, ainsi que les caractéristiques socio-démographiques et l'éducation.

Les différentes variables entre les groupes POP2 et POP3 - dues aux questionnaires différents nous ont permis deux approches exploratoires: l'une orientée vers des informations générales - POP3 - et l'autre avec des informations plus détaillées au sujet des fréquences de consommations alimentaires - POP3 -.

Les variables synthétiques sont ainsi différentes entre ces deux groupes, avec des valeurs communes cependant: Les notions d'éducation et les caractéristiques socio-démographiques par exemple. Mais nous retenons en particulier la malbouffe comme l'un des indicateurs d'IMC élevés, comme cause possible de surpoids et d'obésité.

Annexes

Présentation des données de l'étude INCA3
(Texte produit par l'ANSES avec publication obligatoire)

La 3ème étude Individuelle Nationale des Consommations Alimentaires (INCA3) est une enquête transversale visant à estimer les consommations alimentaires et les comportements en matière d'alimentation des individus vivant en France. L'étude a été menée entre février 2014 et septembre 2015 auprès d'un échantillon représentatif d'individus vivant en France métropolitaine (hors Corse). Au total, 5 855 individus, répartis en 2 698 enfants de la naissance à 17 ans et 3 157 adultes âgés de 18 à 79 ans ont participé à l'étude.

Les individus ont été sélectionnés selon un plan de sondage aléatoire à trois degrés (unités géographiques, logements puis individus), à partir du recensement annuel de la population de 2011, en respectant une stratification géographique (région, taille d'agglomération) afin d'assurer la représentativité sur l'ensemble du territoire. Deux échantillons indépendants ont été constitués : un échantillon « Enfants » (0-17 ans) et un échantillon « Adultes » (18-79 ans).

Les données recueillies dans l'étude portent sur diverses thématiques en lien avec l'évaluation des risques nutritionnels ou sanitaires liés à l'alimentation : consommations d'aliments, de boissons et de compléments alimentaires, habitudes alimentaires (occasions et lieux de consommation, autoconsommation, mode de production des aliments, etc.), pratiques potentiellement à risque au niveau sanitaire (préparation, conservation, consommation de denrées animales crues, etc.), connaissances et comportements en matière d'alimentation. Des données sur les pratiques d'activité physique et de niveau de sédentarité ainsi que sur les caractéristiques socio-démographiques, anthropométriques et de niveau de vie ont également été recueillies.

Afin d'assurer la représentativité nationale des résultats présentés, les individus/ménages participants ont fait l'objet d'un redressement.

Pour les analyses au niveau individuel :

Ce redressement a été réalisé séparément chez les enfants et chez les adultes en tenant compte de variables géographiques et socio-économiques.

Pour les analyses au niveau ménage :

Ce redressement a été réalisé chez l'ensemble des ménages en tenant compte de variables géographiques et socio-économiques.

A chaque individu/ménages est donc associée une pondération prise systématiquement en compte pour les analyses.

Utilisation des données de consommation :

Les consommations alimentaires des individus ont été recueillies sur 3 jours non consécutifs (2 jours de semaine et 1 jour de week-end) répartis sur environ 3 semaines, par la méthode des rappels de 24 heures pour les individus âgés de 15 à 79 ans et par la méthode des enregistrements de 24h (via un carnet alimentaire) pour les individus âgés de 0 à 14 ans. Pour les 3 jours sélectionnés, les individus devaient décrire leurs consommations alimentaires en identifiant tous les aliments et boissons consommés dans la journée et la nuit précédentes. Ils devaient les décrire de façon aussi détaillée que possible et les quantifier à l'aide notamment d'un cahier de photographies de portions alimentaires et de mesures ménagères. Quel que soit l'âge, les interviews étaient conduites par téléphone, à l'aide du logiciel standardisé GloboDiet, par des enquêteurs professionnels spécifiquement formés aux méthodes mises en œuvre et à l'utilisation du logiciel. Parmi les 5 855 individus inclus dans l'étude, 4 114 (2 121 adultes et 1 993 enfants) ont validé le volet consommation en répondant à au moins 2 interviews alimentaires.