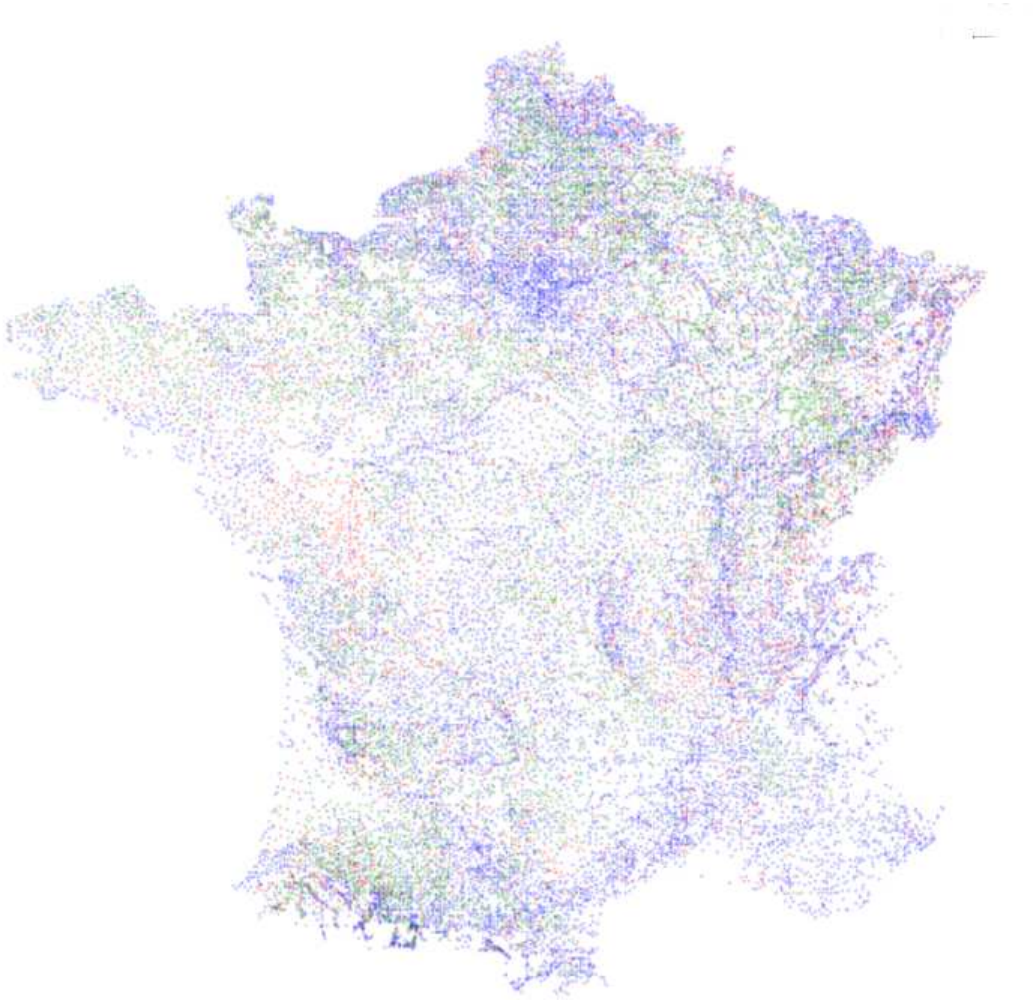
DataScientest • com

Rapport Technique d'évaluation

Projet French Industry

Promotion Juin 2021 - DA Continu

Participants :
Anne Schneider, Omar El Ghazi



1 Table des matières

1 Table des matières 2
1 Contexte 4
1.1 Problématique 4
1.2 Hypothèses 4
2 Data 5
2.1 Analyse du Forme 5
2.1.1 Entreprises 5
2.1.2 Population 7
2.1.3 Salaire Net 8
2.1.4 Géographie 9
2.2 Traitement des données 9
2.2.1 La démarche suivie 9
2.2.2 Ajout et traitement d'un nouveau jeu de données : Consommation annuelle d'électricité et gaz par commune et par code NAF 10
2.2.3 Fusion des jeux de données 11
3 Visualisations 11
3.1 Consommation de gaz et d'électricité par secteur d'activité. 11
3.2 Entreprises 11
3.2.1 Répartition des Entreprises par Région 11
3.2.2 Répartition par taille d'entreprises 13
3.3 Salaire net 14
3.3.1 Répartition du salaire net moyen 14
3.3.2 Salaire net et Population 15
4 Sujets d'analyse 17
4.1 Répartition de la Population (nombre d'habitants par commune) et des Entreprises par secteur d'activité en France Métropolitaine (hors Corse) 17
4.2 Focus sur la Région Rhône-Alpes 21
4.2.1 Répartition de la Population et des Entreprises 21
Dans un rayon de 150km autour de Lyon : 22
4.2.2 Population : répartition par Ages et mode de cohabitation 32
4.2.3 Niveau de Salaire en fonction de la distance à Lyon. 33
4.3 Ecarts de rémunération Femmes / Hommes 34
4.3.1 Ecart dans les catégories socioprofessionnelles 34
4.3.2 Ecart par à l'âge 34
5 Modélisation 35

5.1 Calcul des Clusters sur les communes situées dans un périmètre de 150 km autour de Lyon – sans la variable Salaire 36
5.1.1 Sur le jeu de données Master : Pré-traitement 36
5.1.2 Mise en œuvre de l'Analyse des Composantes Principales (PCA) 39
5.1.3 Mise en œuvre du Modèle de Clustering KMeans 41
5.1.4 Comparaison de la distribution des 2 variables les plus corrélées les 2 composantes principales dans 2 cas de figure 45
5.1.5 Comment les villes, réparties en 4 Clusters sont-elles situées sur une carte ? 47
5.1.6 Conclusion 50
5.2 Calcul des Clusters sur les communes situées dans un périmètre de 140 km autour de Lyon – avec la variable Salaire 51
5.2.1 Pré-traitement 51
5.2.2 Mise en œuvre de l'Analyse des Composantes Principales (PCA) 51
5.2.3 Mise en œuvre du Modèle de Clustering KMeans 54
5.2.4 Visualisation des communes classifiées en 3 clusters, dans un périmètre de 140 km autour de Lyon 56
5.2.5 Analyse des clusters 56
5.2.6 Conclusion 57
6 Bilan 58

1 Contexte

1.1 Problématique

Il est assez intuitif de penser qu'il y existe des disparités entre la France des Grandes Agglomérations, des Grandes Métropoles et celles des territoires en périphérie.
Fernand Braudel décrivait le territoire français comme étant structuré autour d'un système « Villages, Bourgs, Villes ». Il est intéressant de tenter de comprendre les constituants des territoires en périphérie, entre deux Grandes Villes ou deux Métropoles.
Quel est le rapport entre le tissu économique et la répartition de la population, sur les territoires compris entre les grandes agglomérations ? Quelles sont les disparités entre les territoires ? Comprendre la répartition de la population, des salaires et des entreprises par secteur d'activité, sur les différents territoires en France Métropolitaine, et plus particulièrement en région.

1.2 Hypothèses

Nos hypothèses de travail sont que :

- Dans les grandes villes et les grandes agglomérations, et à leur proche périphérie, on trouve principalement des entreprises du secteur tertiaire, une "forte" densité de population et un revenu moyen élevé.
- Et en s'éloignant de plus en plus des grandes agglomérations on trouve principalement des entreprises du secteur industriel, puis agricole, une densité de population moindre et des revenus moins élevés, jusqu'à se rapprocher d'une autre grande agglomération.

Critères d'évaluation :

- Taille de la population
- Ages prédominants
- Niveau de revenus
- Secteur d'activité des Entreprises.

On cherchera donc si, parallèlement aux critères démographiques et économiques communément utilisés pour décrire les disparités entre les différents territoires, les secteurs d'activité prépondérants sont également différents.

On utilisera les consommations en gaz et en électricité de chacun des secteurs, Agricole, Industriel, Tertiaire. Le secteur d'activité avec la plus forte consommation en gaz et électricité sera considéré comme le secteur d'activité prépondérant de la commune. On identifiera avec cette méthode les communes avec une activité principalement Agricole, Industrielle ou Tertiaire.

Fernand Braudel : L'identité de la France. Espace et histoire, Paris, Arthaud / Flammarion, 1986

2 Data

2.1 Analyse du Forme

2.1.1 Entreprises

A) Chargement et contenu du dataset "Entreprises"
base_etablissements_par_tranches_effectif

➔ La liste des variables :

VAR_ID	VAR_LIB_LONG
CODGEO	Code de la zone géographique
LIBGEO	Libellé de la zone géographique
reg	Code Région
dep	Code Département
E14tst	Nombre Total d'Entreprises dans la commune
E14ts0ND	Nombre d'Entreprises avec 0 salarié ou un nombre de salarié non déterminé
E14ts1	Nombre d'Entreprises avec 1 à 5 salariés
E14ts6	Nombre d'Entreprises avec 6 à 10 salariés
E14ts10	Nombre d'Entreprises avec 11 à 20 salariés
E14ts20	Nombre d'Entreprises avec 21 à 50 salariés
E14ts50	Nombre d'Entreprises avec 51 à 99 salariés
E14ts100	Nombre d'Entreprises avec 100 à 199 salariés
E14ts200	Nombre d'Entreprises avec 200 à 499 salariés
E14ts500	Nombre d'Entreprises avec 500 salariés et plus

Ce jeu de données nous informe sur le nombre d'entreprises dans chacune des 36681 communes de France.
36681 entrées (communes) et 14 colonnes correspondant à :

⇨ **Les informations sur la commune**

La variable CODGEO correspond au code INSEE de la commune : identifiant unique pour chaque commune en France
La variable LIBGEO correspond au nom de la commune, qui n'est pas forcément unique.

La variable REG correspond au code de la région dans laquelle est située la colonne

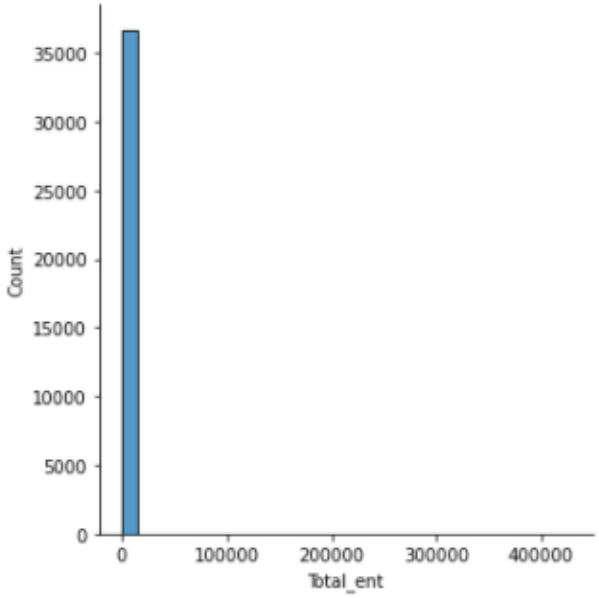
On constate 27 codes régions. Alors que depuis janvier 2016, la France est découpée en seulement 18 régions (13 Métropolitaines et 5 ultra-marines (https://www.data.gouv.fr/fr/datasets/carte-des-regions-francaises-nouvelles-regions-de-2016/.
Il sera certainement intéressant d'ajuster les Régions sur cette nouvelle taxonomie

La variable DEP correspond au Département dans lequel est situé la commune
Donc, LIBGEO, DEP et REG ont une dépendance hiérarchique.

=> Les informations relatives au nombre d'entreprises dans la commune

La variable E14TST nous informe du nombre total d'entreprises dans la commune, les autres variables quantitatives représentent la répartition de ce nombre d'entreprise par taille d'entreprise.

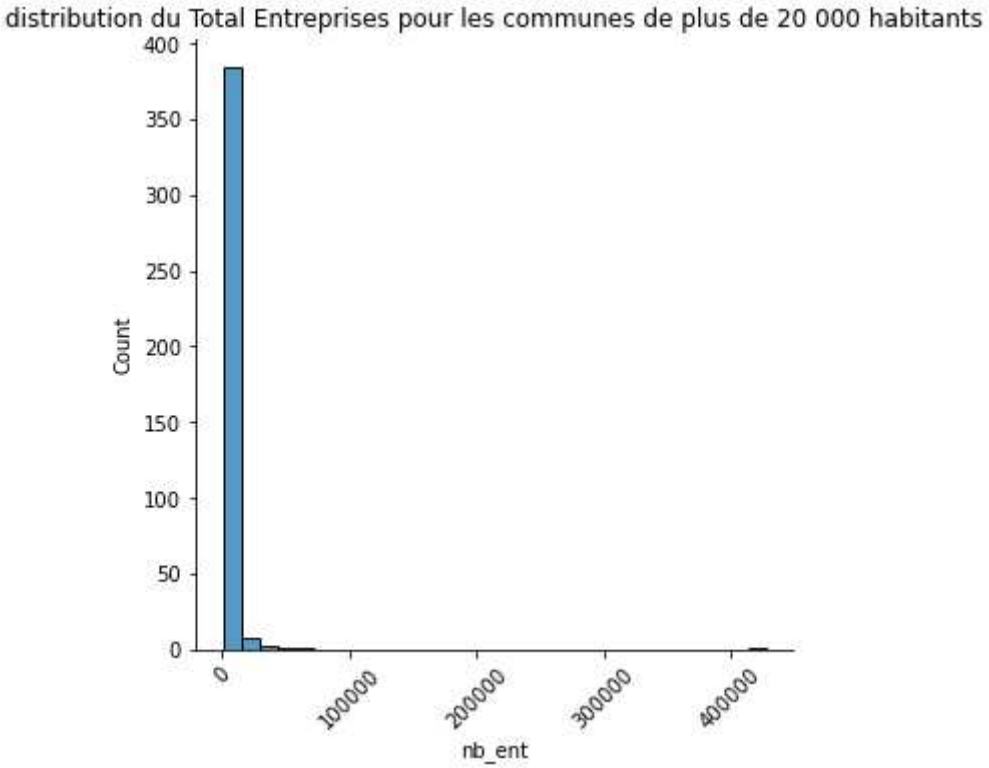
B) Analyse du Forme : Distribution et équilibre des variables quantitatives :
Seule la variable principale E14TST (Total_ent) sera utilisée, les autres variables étant dérivées de cette dernière :



Le graphique illustre bien le déséquilibre de la variable. On constate que la très grande majorité des villes compte moins de 25 000 entreprises par commune.

Avec sans surprise, les plus grandes communes de France dans le TOP5, en termes de nombre d'entreprises par communes :

	CODGEO	LIBGEO	REG	DEP	Total_ent
30784	75056	Paris	11	75	427385
4453	13055	Marseille	93	13	68332
28522	69123	Lyon	82	69	49756
2014	06088	Nice	93	06	39314
12418	31555	Toulouse	73	31	36823



Pour les communes de plus de 20 000 habitants, on note que la grande majorité des communes comprennent sur leur territoire un nombre d'entreprises inférieur à 20 000, et seule une commune (Paris) comprends plus de 400 000 entreprises. Soit une très grande disparité entre Paris et les autres grandes communes de France.

2.1.2 Population

⇨ Liste des variables

VAR_ID	VAR_LIB_LONG
NIVGEO	Niveau géographique (arrondissement, communes…)
CODGEO	Code de la zone géographique
LIBGEO	Libellé de la zone géographique
MOCO	mode de cohabitation : 11 = enfants vivant avec deux parents 12 = enfants vivant avec un seul parent 21 = adultes vivant en couple sans enfant 22 = adultes vivant en couple avec enfants 23 = adultes vivant seuls avec enfants 31 = personnes étrangères à la famille vivant au foyer 32 = personnes vivant seules
AGE80_17	catégorie d'âge (tranche de 5 ans) | ex : 0 -> personnes âgées de 0 à 4 ans
SEXE	sexe, 1 pour homme | 2 pour femme
NB	Nombre de personnes dans la catégorie

Ce jeu de données contient 8536584 lignes et 7 colonnes. Les 36 000 communes de France sont représentées.

2.1.3 Salaire Net

⇨ Liste des variables

VAR_ID	VAR_LIB_LONG
CODGEO	Code de la zone géographique
LIBGEO	Libellé de la zone géographique
SNHM14	Salaire net horaire moyen en 2014 (€)
SNHMC14	Salaire net horaire moyen des cadres, professions intellectuelles supérieures et des chefs d'entreprises salariés en 2014 (€)
SNHMP14	Salaire net horaire moyen des professions intermédiaires en 2014 (€)
SNHME14	Salaire net horaire moyen des employées en 2014 (€)
SNHMO14	Salaire net horaire moyen des ouvriers en 2014 (€)
SNHMF14	Salaire net horaire moyen des femmes en 2014 (€)
SNHMFC14	Salaire net horaire moyen des femmes cadres, professions intellectuelles supérieures et des chefs d'entreprises salariés en 2014 (€)
SNHMFP14	Salaire net horaire moyen des femmes exerçant une profession intermédiaire en 2014 (€)
SNHMFE14	Salaire net horaire moyen des femmes employées en 2014 (€)
SNHMFO14	Salaire net horaire moyen des femmes ouvrières en 2014 (€)
SNHMH14	Salaire net horaire moyen des hommes en 2014 (€)
SNHMHC14	Salaire net horaire moyen des hommes cadres, professions intellectuelles supérieures et des chefs d'entreprises salariés en 2014 (€)
SNHMHP14	Salaire net horaire moyen des hommes exerçant une profession intermédiaire en 2014 (€)
SNHMHE14	Salaire net horaire moyen des hommes employés en 2014 (€)
SNHMHO14	Salaire net horaire moyen des hommes ouvriers en 2014 (€)
SNHM1814	Salaire net horaire moyen des personnes de 18 à 25 ans en 2014 (€)
SNHM2614	Salaire net horaire moyen des personnes de 26 à 50 ans en 2014 (€)
SNHM5014	Salaire net horaire moyen des personnes de plus de 50 ans en 2014 (€)
SNHMF1814	Salaire net horaire moyen des femmes de 18 à 25 ans en 2014 (€)
SNHMF2614	Salaire net horaire moyen des femmes de 26 à 50 ans en 2014 (€)
SNHMF5014	Salaire net horaire moyen des femmes plus de 50 ans en 2014 (€)
SNHMH1814	Salaire net horaire moyen des hommes de 18 à 25 ans en 2014 (€)
SNHMH2614	Salaire net horaire moyen des hommes de 26 à 50 ans en 2014 (€)
SNHMH5014	Salaire net horaire moyen des hommes de plus de 50 ans en 2014 (€)

Ce jeu de données contient 5136 lignes et 26 colonnes sans aucune valeur manquante.
A l'exception de CODGEO et LIBGEO, toutes les autres variables sont numériques.

La variable SNHM14 indique le salaire net horaire moyen (€) constaté pour chaque commune en 2014. Les autres variables quantitatives représentent la répartition de cette variable par les différentes catégories socioprofessionnelles ou sa répartition par groupes d'âge.

⇨ Nombre de communes

Il était important de vérifier pourquoi nous disposons que d'un nombre réduit de communes sur ce jeu de données comparant aux autres jeux de données.

Nous avons donc cherché la réponse et les explications à partir des informations disponibles sur le site de l'INSEE.

https://www.insee.fr/fr/statistiques/2021266

Selon l'INSEE :
Les données issues des Bases Tous Salariés sont soumises au secret statistique. Aucune statistique n'est diffusée pour les zones de moins de 2 000 habitants. Pour quelques zones de 2 000 habitants ou plus, qui ne respectent pas les seuils, il n'est pas non plus possible de diffuser des résultats. À savoir, chaque case du tableau doit comporter au moins 5 salariés et aucun salarié ne doit représenter plus de 80% de la masse salariale de la case.

D'autre informations communiquées qui sont utiles pour notre analyse :

Les statistiques sont établies à partir des informations recueillies sur les entreprises du secteur privé et les entreprises publiques localisées en France.
Les statistiques sur les catégories socioprofessionnelles portent sur le poste principal occupé par le salarié dans l'année, hors agriculture et catégorie socioprofessionnelle non définie.
Les personnes dont l'âge n'est pas renseigné et les mineurs sont également exclus du champ statistique.

⇨ Calcul du salaire horaire net moyen
Voici comment il est défini et calculé par l'INSEE :
Résultat du quotient de la masse des salaires nets rapportée au nombre d'heures salariés calculé sur tous les postes effectués par le salarié au cours de l'année (hors indemnités chômage). Le nombre d'heures salariées prend en compte les heures supplémentaires rémunérées et toutes les périodes au cours desquelles le salarié demeure lié à un établissement du fait du contrat de travail (congés, période de maladie et d'accident de travail), à l'exception des périodes de congés sans solde.

2.1.4 Géographie

Chargement et contenu du jeu de donnée "Géographie"

Il est le jeu de donné commun à l'ensemble des jeux disponibles. Dans le sens où tous seront fusionnés à Géographie.
Ce jeu de données sera principalement utilisé pour compléter les informations géographiques sur les communes, ainsi que les données de longitude et latitudes disponible, mais par pour la totalité des communes, et ce afin de calculer des distances entre des communes, ou des agglomérations. Aussi il ne sera pas analysé.

2.2 Traitement des données

2.2.1 La démarche suivie

Les traitements suivants ont été appliqués sur l'ensemble des jeux de sonnées.

- Traitement CODGEO ou code commune

Cette information est la clé commune dans tous nos jeux de données et elle correspond au code INSEE de la commune qui est présente dans le jeu de données Géographie.

Cependant ces deux informations sont différentes pour les communes de la Corse et celles des DOM TOM. Par conséquence, la Corse et les DOM TOM feront objet de suppression dans tous nos jeux de données. Nous ne disposons pas de données pour Monaco.

Cela est sans impact sur l'analyse que nous souhaitons réaliser.
D'une part pour une raison technique, car nous ne disposons pas des données de longitude et latitude dans le jeu de données, pour les DOM TOM, et d'autre part, car nous supposons une spécificité territoriale pour ces territoires insulaires (Dynamiques spécifiques dues à leur caractère insulaire, voire leur distance avec la Métropole pour les DOM TOM.

- **Normalisation des colonnes**
- **Suppression des colonnes inutiles**
- **Retraitement des villes à arrondissement**
- **Suppression ou substitution des valeurs manquantes**
- **Sauvegarde des jeux de données nettoyés dans le répertoire "clean_data**"

2.2.2 Ajout et traitement d'un nouveau jeu de données : Consommation annuelle d'électricité et gaz par commune et par code NAF

Ce jeu de données, à la base, permet de visualiser l'évolution année par année (de 2011 à 2019) des consommations d'électricité et de gaz en MWh et du nombre de points de livraison, par secteur (résidentiel, tertiaire, industriel, agricole ou non affecté), par catégorie de consommation, par code NAF et par maille géographique.

Cependant, nous exploiterons, dans un premier temps, ce jeu de données pour identifier en consommation moyenne la part des grands secteurs d'activité (Résidentiel, Tertiaire, Industrie, Agriculture ou Secteur Inconnu) pour chaque commune.

On utilisera les consommations en gaz et en électricité de chacun des secteurs, Agricole, Industriel, Tertiaire. Le secteur d'activité avec la plus forte consommation en gaz et électricité sera considéré comme le secteur d'activité prépondérant de la commune. On identifiera avec cette méthode les communes avec une activité principalement Agricole, Industrielle ou Tertiaire.

Comme il nous est impossible d'analyser la variable activité à travers les précédents jeux de données, l'introduction de ce jeu de donnée est donc justifiée et nécessaire pour que nous puissions vérifier notre hypothèse.

Ce nouveau jeu de données a fait objet à la même démarche du retraitement que les jeux de données précédents :

2.2.3 Fusion des jeux de données

Comme notre jeu de données "salary_clean" ne contient que 5023 lignes, nous avons choisi de traiter deux versions de fusion : avec et sans salaire net. Cela nous permettrait de ne pas réduire le champ d'analyse sur les activités, population et industrie pour lesquelles la fusion sans salaire est plus pertinente.

2.2.3.1 Fusion sans salaire

- Le jeu de données "geography" a été utilisé comme base pour fusionner les jeu de données "entreprises", "population" et "consommation"
- Le fichier issu de la fusion et est nommé master et sauvegardé dans "clean_data"

2.2.3.2 Fusion avec Salaire

- Le fichier salaire a été utilisé comme base pour fusionner le fichier master issu de la fusion sans salaire.
- Le fichier issu de la fusion avec salaire salary_master est nommé master et sauvegardé dans "clean_data"

3 Visualisations

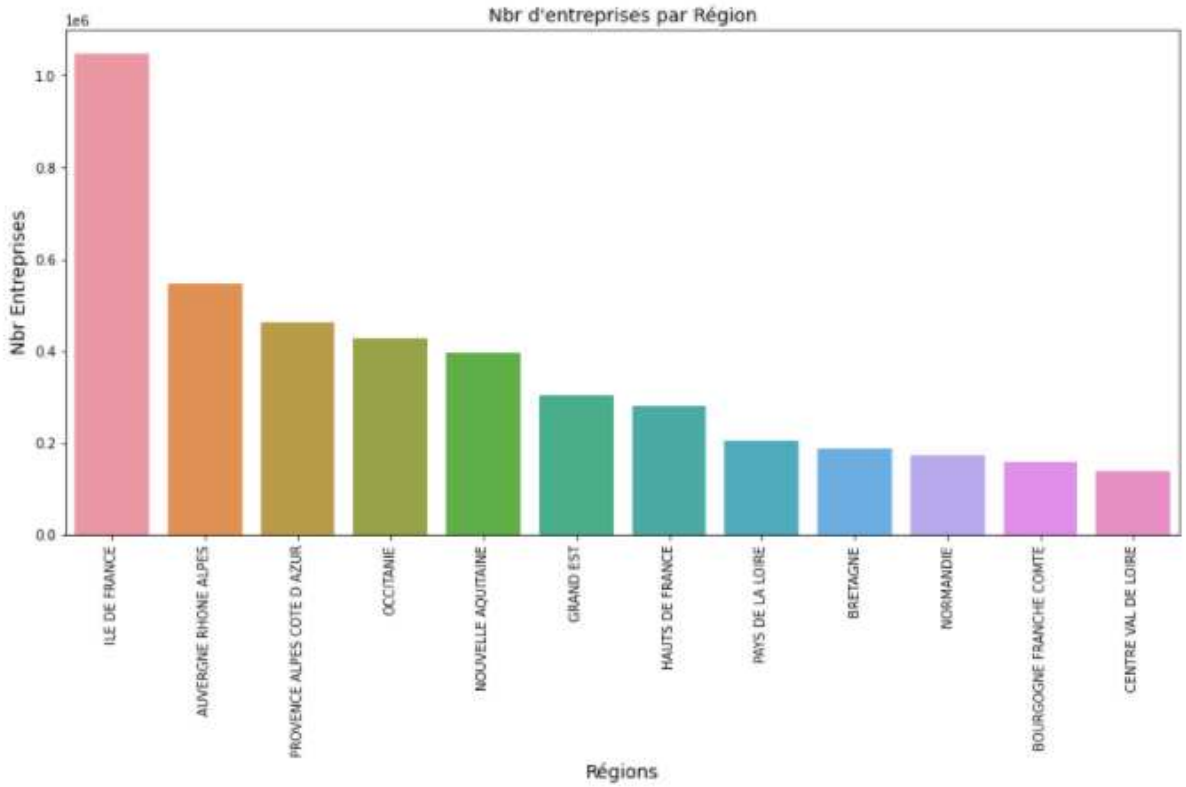
3.1 Consommation de gaz et d'électricité par secteur d'activité.

Comme pour géographie – pas de visualisation à faire car on utilise ce jeu de données pour récupérer l'activité des entreprises et la prédominance par commune en fonction de la consommation par secteur d'activité dans la commune.

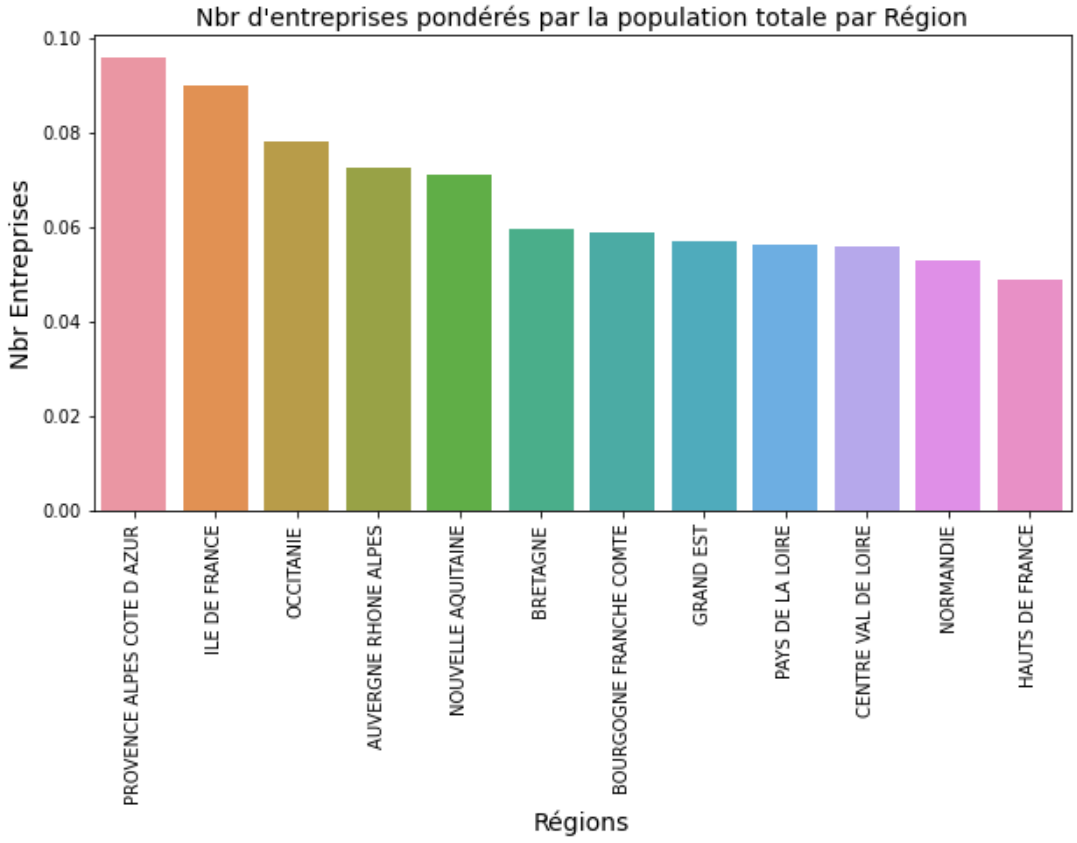
3.2 Entreprises

3.2.1 Répartition des Entreprises par Région

Du point de vue graphique, il semble pertinent de mesurer le nombre d'entreprises par Région (sur les 18 Régions, plutôt que par communes). Le jeu de données initiale fournit une information en tenant compte de l'ancienne classification sur 27 Régions.
L'INSEE fournit un fichier comprenant les codes des anciennes régions, correspondant aux codes des 18 nouvelles régions et leur libellé. La fusion des deux jeux permet de produire des graphiques suivants :



Sans surprise, les Régions comportant le plus d'entreprises en France sont l'Ile de France (24% des Entreprises du territoire Métropolitain hors Corse), la région Rhône-Alpes (13%), Provence Alpes Côte d'Azur, l'Occitanie et la Région Nouvelle Aquitaine.



En pondérant le nombre total d'entreprises par la population totale de la commune, on identifie que la Région PACA est celle comportant le plus d'entreprises par habitants, la Région Ile de France n'arrivant qu'en deuxième place.

3.2.2 Répartition par taille d'entreprises

Quel est la taille d'entreprise la plus représentée en France ?
Le jeu de données répartit les entreprises par selon leur taille, en 9 catégories.
L'Union Européenne propose une classification des entreprises selon leur effectif, en 4 catégories,(https://ec.europa.eu/eurostat/statisticsexplained/index.php?title=Glossary:Enterprise_size

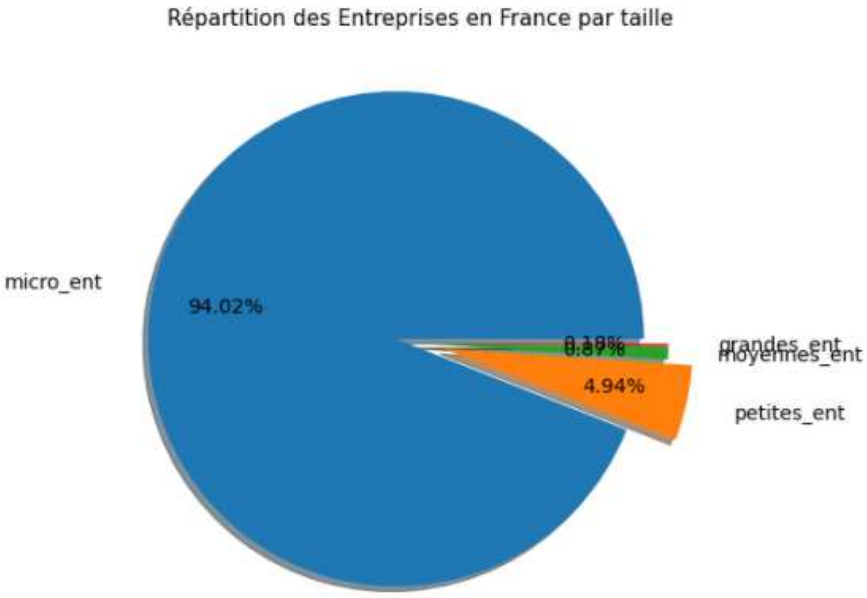
- **small and medium-sized enterprises**, abbreviated as **SMEs**: *fewer than 250* persons employed:

  SMEs are further subdivided into:

  - **micro enterprises**: *fewer than 10* persons employed;
  - **small enterprises**: *10 to 49* persons employed;
  - **medium-sized enterprises**: *50 to 249* persons employed;

- **large enterprises**: *250 or more* persons employed.

Cette classification a été appliquée au jeu de données afin de visualiser la proportion des entreprises en France en fonction de leur effectif :

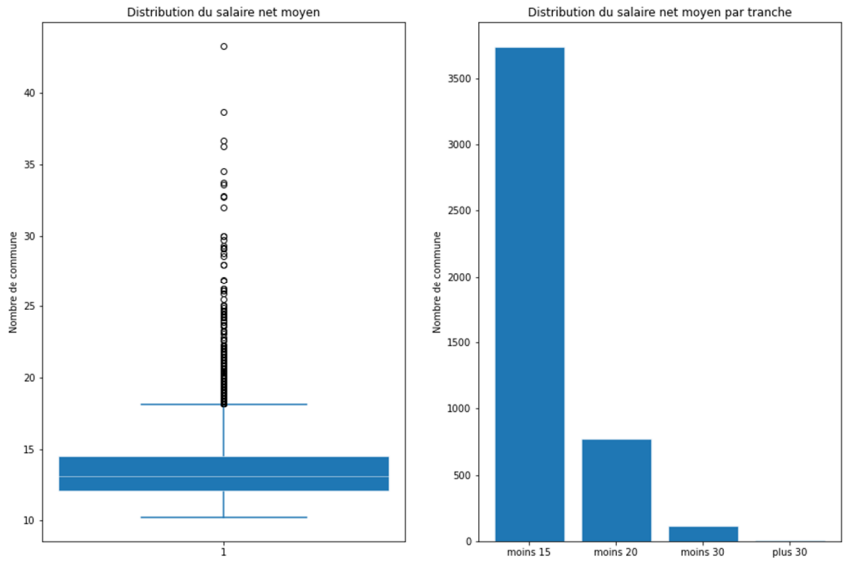


En France, les entreprises sont à 99% des micro et petites entreprises de 0 à 10 salariés. La répartition est la même au niveau Régional.

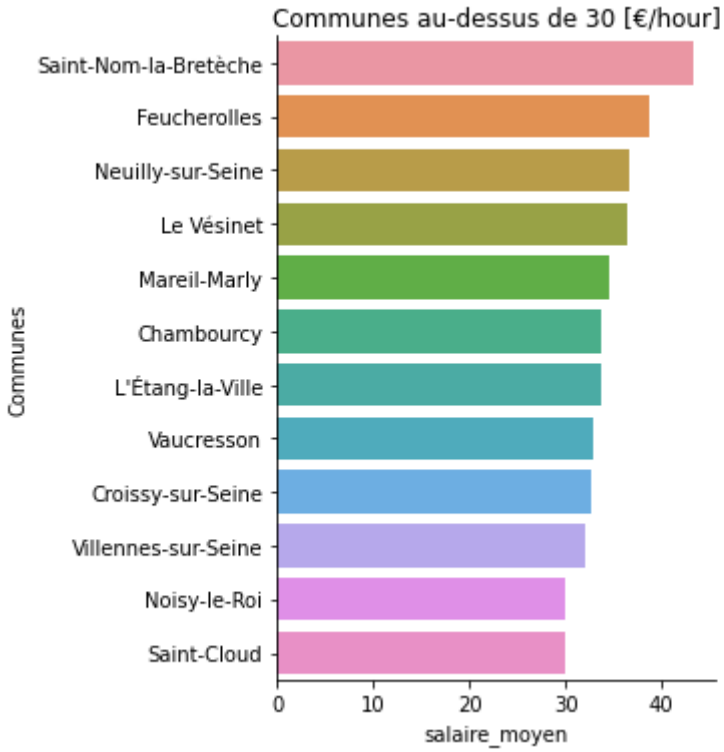
3.3 Salaire net

3.3.1 Répartition du salaire net moyen

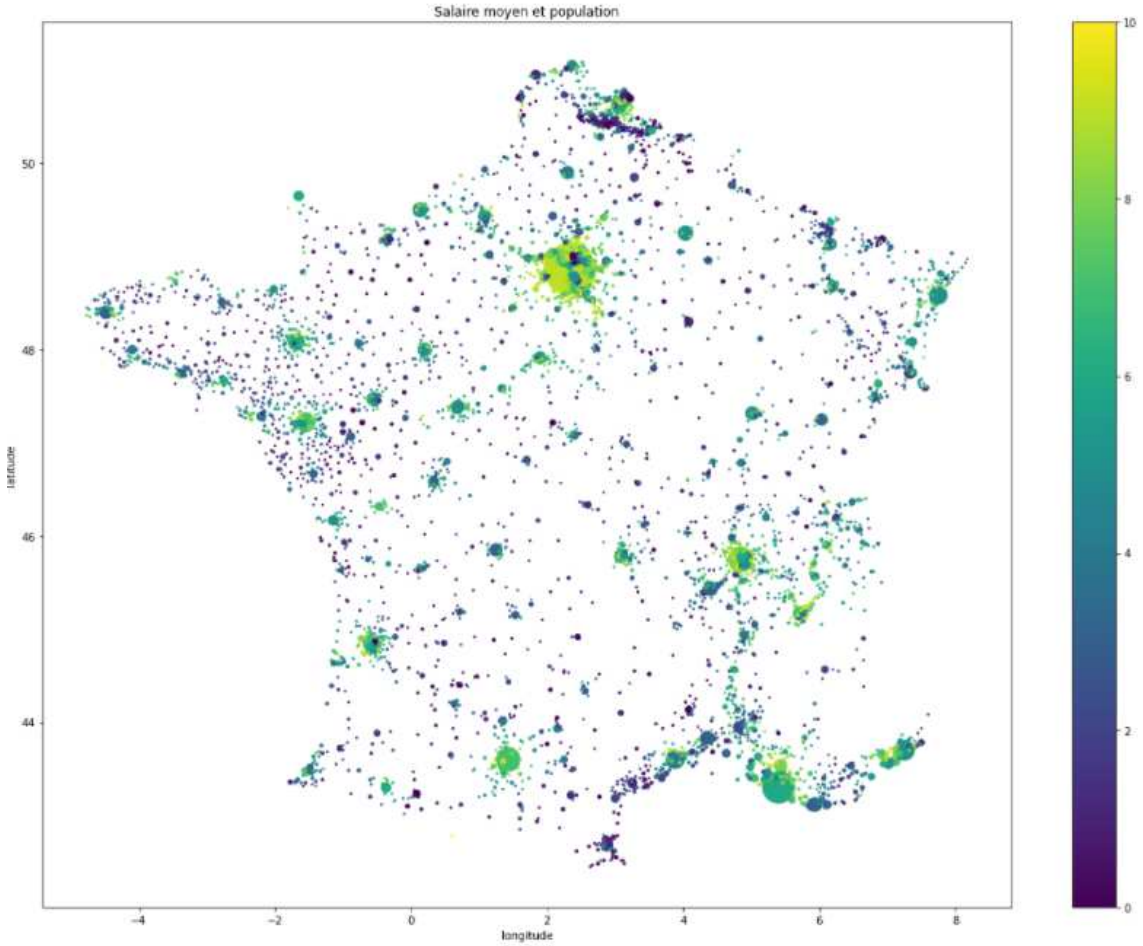
Le salaire moyen semble bien être similaire dans la majorité des communes de France
Néanmoins, il y a bien un groupe de communes dont le salaire est très élevé par rapport à la moyenne nationale.



Sans surprise les communes les plus riches se situent dans les départements 78 et 92.

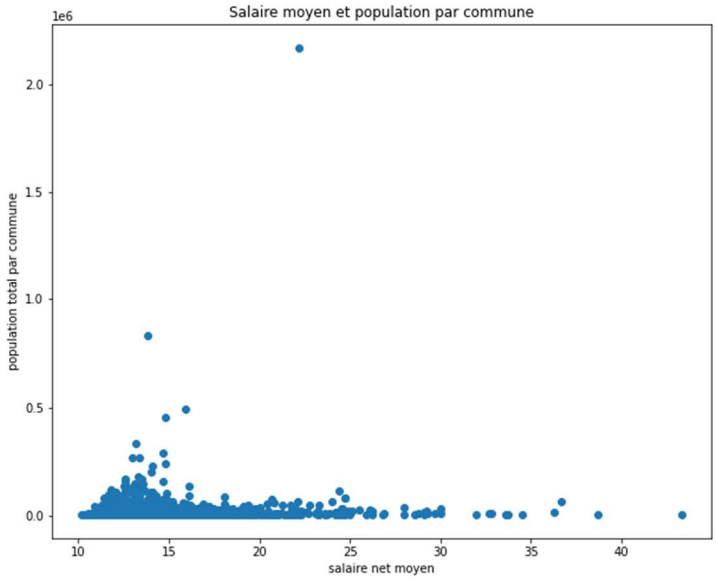


3.3.2 Salaire net et Population



La taille des points sur la carte est proportionnelle au nombre d'habitants dans les communes, plus la taille du point est grande plus la population est importante.
La couleur des points sur la carte indique le niveau du salaire moyen, plus la couleur est claire plus le salaire moyen est élevé.
Il semble que les points de couleurs plus claires (salaire net élevé) se situent dans les zones ayant des populations importantes.

Analysons si il y a un lien de corrélation entre le nombre d'habitants dans une commune et le salaire net moyen.



Il semble qu'il n'y a pas de corrélation entre la taille de la population et le salaire net moyen pour les communes en France.

Cela nous permet donc de supposer que les agglomérations avec des populations importantes renforcent le salaire net moyen (les points de couleurs claires sur la carte).

4 Sujets d’analyse

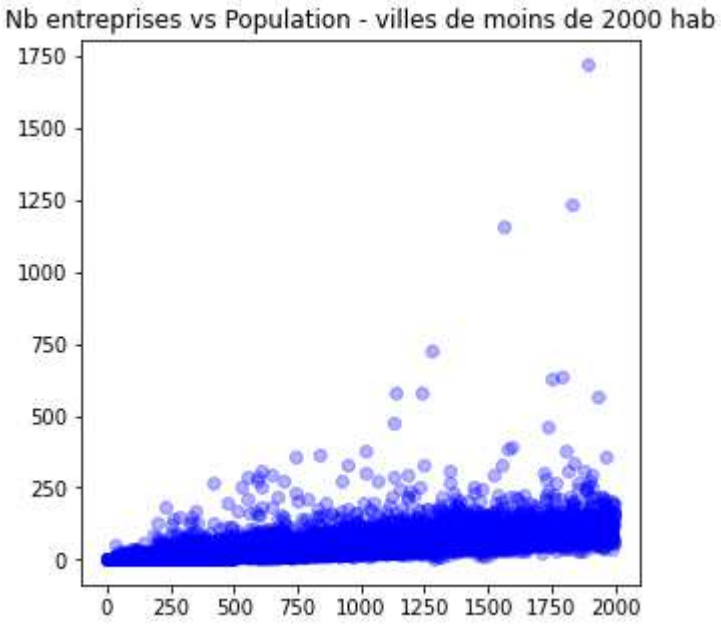
Dans les grandes villes et les grandes agglomérations, et à leur proche périphérie, on trouve principalement des entreprises du secteur tertiaire, une “forte” de population et un revenu moyen élevé. Et en s’éloignant de plus en plus des grandes agglomérations on trouve principalement des entreprises du secteur industriel, puis agricole ou artisanal, une densité de population moindre et des revenus moins élevés, jusqu’à se rapprocher d’une autre grande agglomération.

Critères d’évaluation :

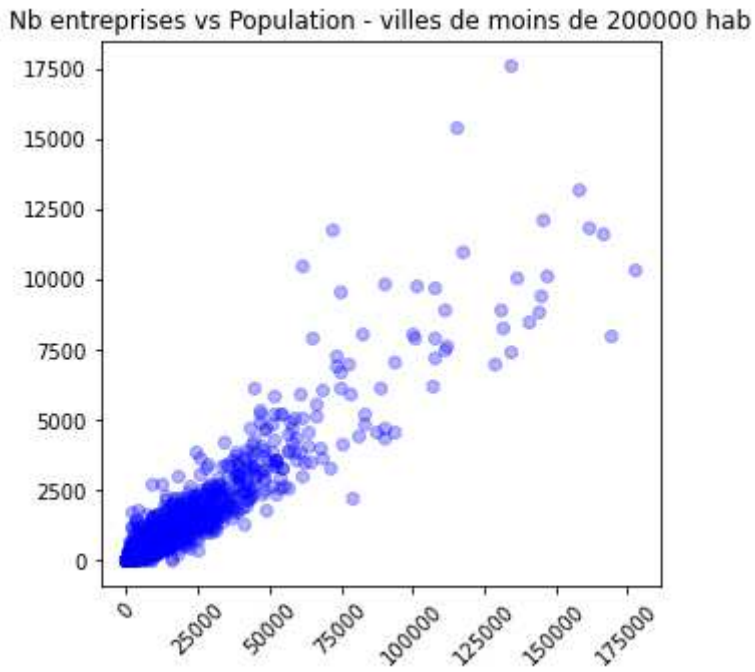
- Taille de la population
- Secteur d’activité des Entreprise et Nombre d’entreprises par secteur
- Ages prédominants
- Niveau de revenus, différence Femmes / Hommes

4.1 Répartition de la Population (nombre d’habitants par commune) et des Entreprises par secteur d’activité en France Métropolitaine (hors Corse)

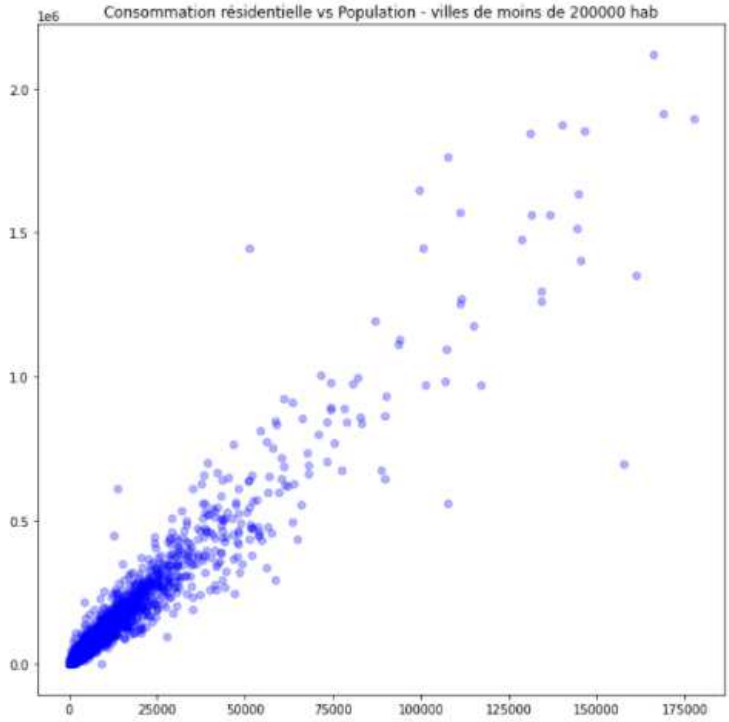
a. Relation entre la taille de la population et le nombre d’entreprises dans les communes



80% des communes en France ont une population inférieure à 2000 habitants.
Pour la très grande majorité de ces communes, le nombre d’entreprises sur la commune est inférieur à 250.
En considérant les communes de moins de 200 000 habitants, on note une relation linéaire entre la taille de la population et le nombre d’entreprises sur la commune.

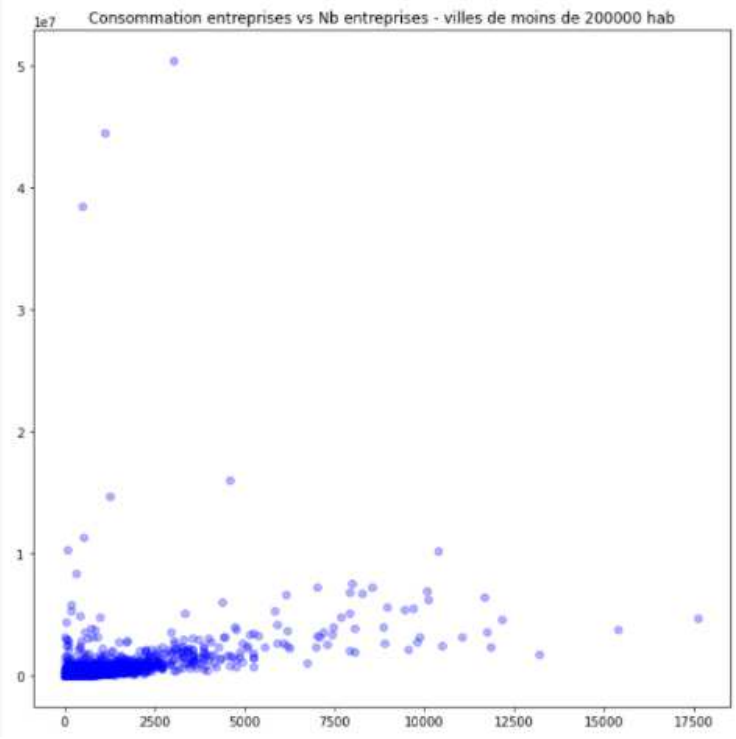


Consommation résidentielle des communes en fonction du nombre d'habitants



Les données semblent cohérentes, car assez logiquement, la relation est croissante et linéaire

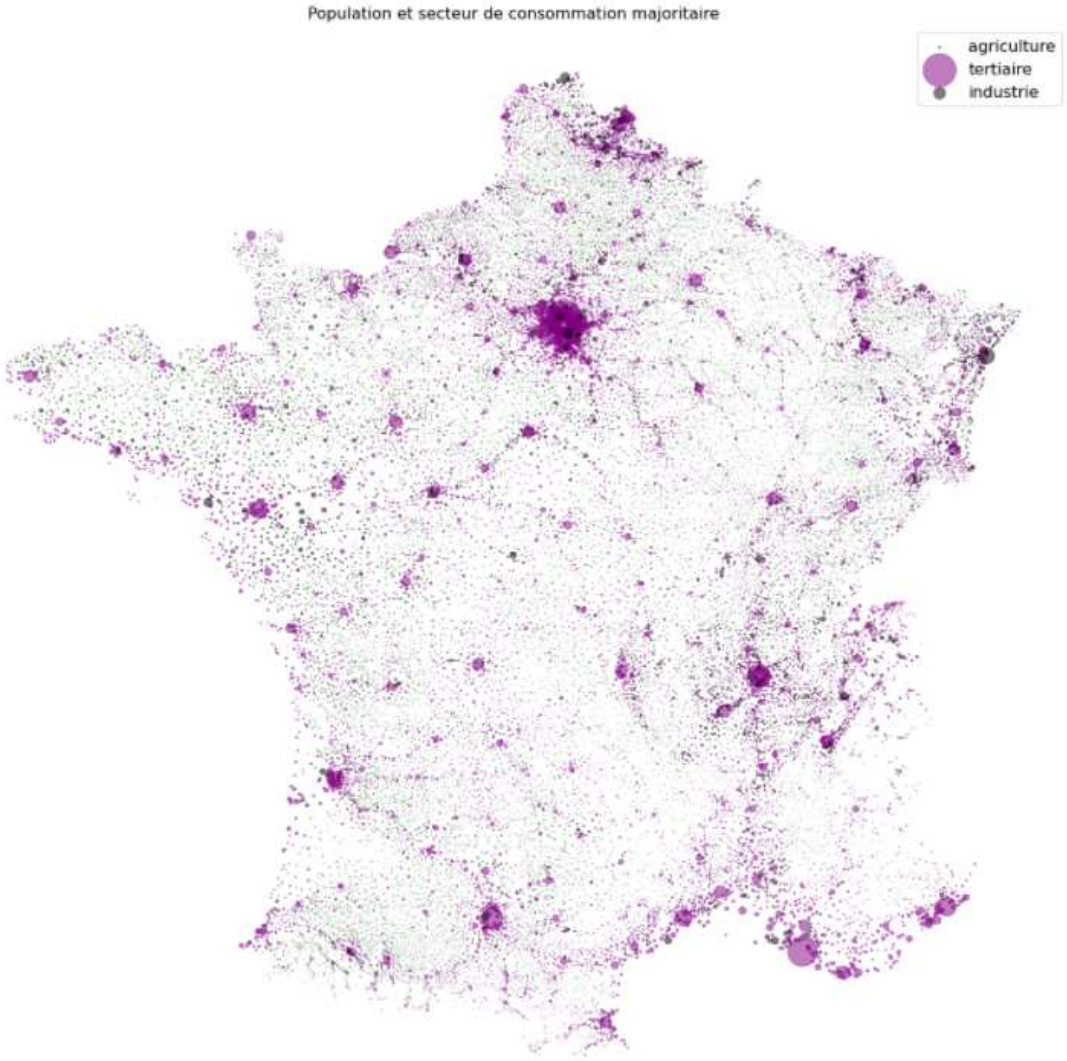
Consommation des entreprises en fonction du nombre d'entreprises dans la commune



La relation est plus compliquée à expliquer. Naturellement, il paraît logique que la consommation en gaz ou en électricité dépende de l'activité de l'entreprise.
En effet, les communes de ce graphique présentant les plus fortes consommations d'énergie comptent sur leur territoire des activités de Fonderie ou de Raffinerie :

	nom_commune	nom_region	nb_ent	consommation_libelle_categorie_consommation_entreprises
17545	Blénod-lès-Pont-à-Mousson	Lorraine	162.0	5.851436e+06
27031	Notre-Dame-de-Gravenchon	Haute-Normandie	324.0	8.351110e+06
24208	Bantzenheim	Alsace	57.0	1.032321e+07
14631	Montoir-de-Bretagne	Pays de la Loire	507.0	1.139164e+07
26850	Gonfreville-l'Orcher	Haute-Normandie	479.0	3.852548e+07

b. Répartition sur le territoire national de la Population et du secteur de consommation majoritaire



La taille du point dépend du nombre d'habitants. La couleur du point dépend du secteur de consommation prépondérant sur la commune. Par exemple, la consommation en gaz et électricité du secteur Tertiaire plus élevée que celle des secteurs industriels ou agricoles sur la commune, on considère alors que l'activité prépondérante de la commune est Tertiaire.

Les zones autour de Paris, Marseille, Toulouse, Bordeaux Nantes et Lille sont bien sûr les secteurs les plus peuplés, et le secteur d'activité Tertiaire est prépondérant.
Et de façon plus générale, le secteur d'activité prépondérant sur les communes les plus peuplées semble être le secteur tertiaire (finance, immobilier, transport, tourisme, administration, éducation...), les zones périphériques à ces dernières sont moins peuplées et semblent avoir une activité prépondérante agricole. (Phénomène assez marqué en Bretagne, dans les terres par exemple, Rennes et les villes côtières principales étant les plus peuplées.).

On remarque peu de « Grandes Villes » Industrielles, à part dans le secteur de Strasbourg , de l'Ile de France, dans le Nord.
Ceci peut être corroboré avec la « tertiairisation » de la France décrite par Jérôme Fourquet dans son dernier ouvrage : « la France devant nos Yeux ».

« La France sous nos Yeux » Jérôme Fourquet, Jean-Laurent Cassely, Seuil, Oct. 2021

4.2 Focus sur la Région Rhône-Alpes

La Région Rhône-Alpes a une situation particulière sur le territoire français et même européen.
En effet, Lyon et sa région sont sur l'axe Paris – Marseille, et la région est aussi à proximité de la frontière avec l'Italie, la Suisse, et est également le carrefour de plusieurs axes de transport trans européens (Autoroute A6, Ligne TGV, et 2 gares TGV, et le quatrième aéroport de France : Lyon Saint-Exupéry).

La région Rhône-Alpes est également assez connue pour les activités Industrielles (industrie chimique, pharmaceutique, automobile et métallurgie).

Tout en étant une région agricole également reconnue (Viticulture – des Côtes-Rôties au sud jusqu'à la Bourgogne au Nord), et enfin pour ses espaces naturels (les Alpes, la proximité plusieurs Parcs Naturels Régionaux – Livradois-Forez, le Vercors…).

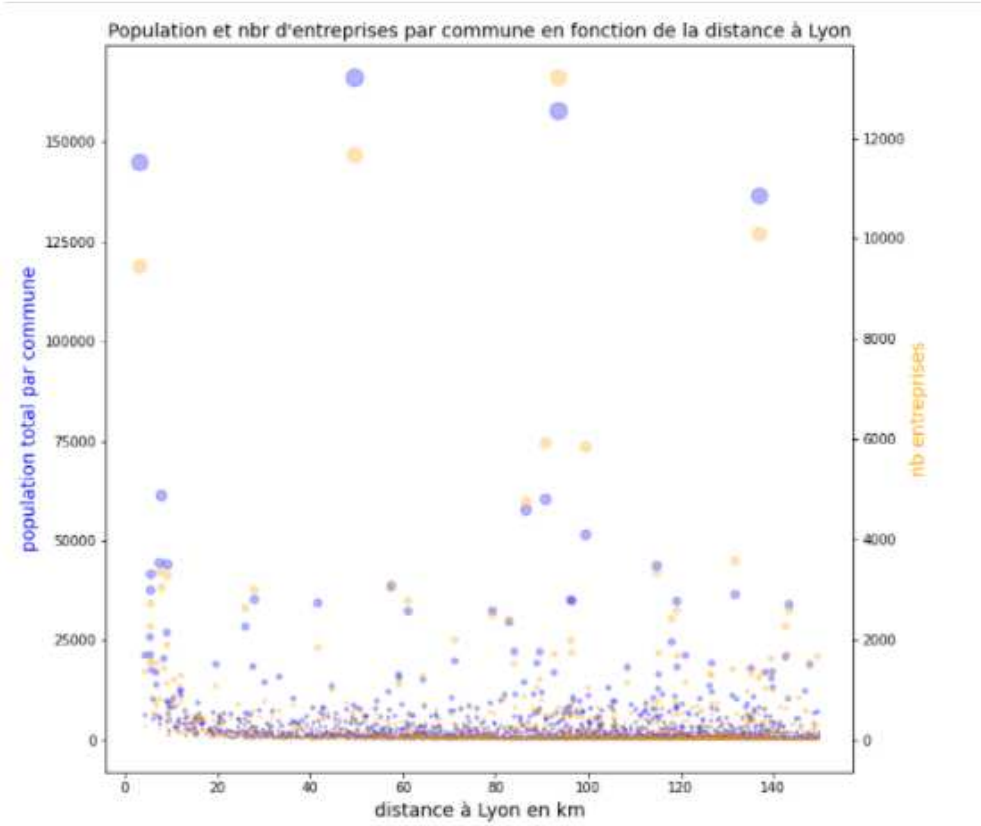
La proximité avec plusieurs grandes villes, comme Grenoble, Saint-Etienne, Valence, Clermont-Ferrand, at aussi Genève (Suisse), justifie de choisir un périmètre assez large autour de Lyon afin de visualiser le possible effet polarisant de ces villes importantes sur les communes périphériques, en plus de l'effet « Lyon », et de l'effet de la distance entre la Métropole et les communes en périphérie.

4.2.1 Répartition de la Population et des Entreprises

En fonction de la distance à la Métropole Lyon, comment se répartissent les entreprises par secteur d'activité et la population ?

4.2.1.1 Répartition Population et Entreprises

Dans un rayon de 150km autour de Lyon :



Ce graphique est riche en informations :

- On constate qu'à moins de 3km des coordonnées GPS données pour Lyon on se situe encore sur la ville de Lyon
- Une ville à 50 km de Lyon a une population supérieure à 150 000 habitants. Il s'agit de **Saint-Etienne.**
- A proximité de Lyon, les communes ont une population et un nombre plus important d'entreprises que lorsqu'on s'éloigne de Lyon.
- A partir d'un rayon situé à 30 km de Lyon, et tous les 20 km, on trouve des communes dont la population est comprise entre 25 000 et 50 000 habitants

Les cartes de transports régionaux et autoroutes montrent que ces communes ont des "nœuds" où plusieurs lignes ferroviaires et réseaux de route se croisent. *(cf. cartes p42)*

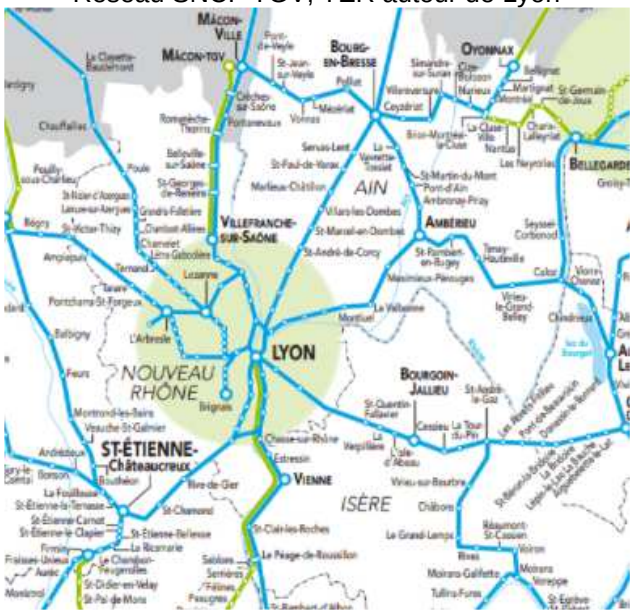
- **Enfin, au-delà de 90 km de Lyon, on constate à nouveau un nombre important de communes de plus de 50 000 habitants, et comptant plus de 2000 entreprises. Il s'agit de Métropoles : Grenoble, Clermont-Ferrand, Valence. (***Tableau en bas de page)*

Aussi, pour la suite de l'analyse on considèrera les communes situées à 150 km de Lyon.

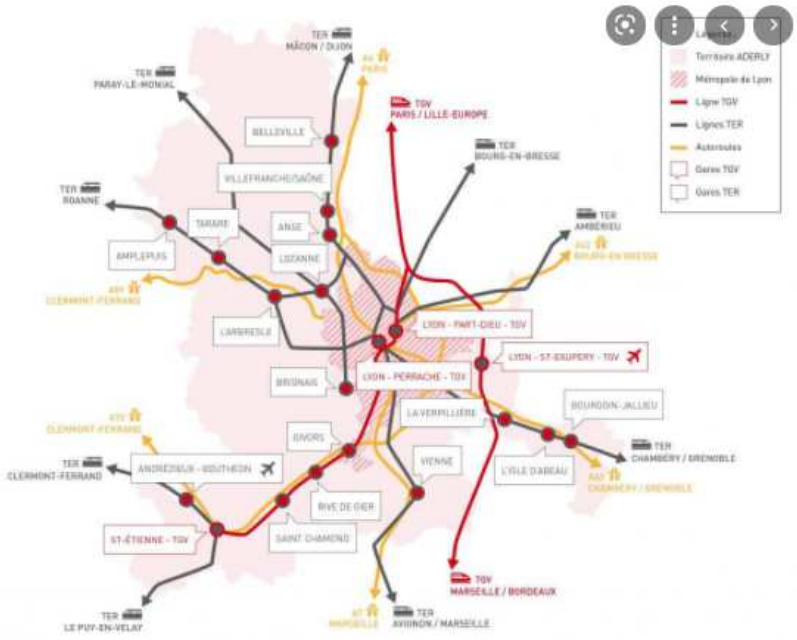
Tableau : Villes de plus de 50 000 habitants, et situées à plus de 20 km de Lyon.

	nom_commune	pop_total	nb_ent	distance_to_commune
14197	Saint-Etienne	166137.0	11663.0	49.646994
12650	Grenoble	157802.0	13207.0	93.661702
22224	Clermont-Ferrand	136606.0	10090.0	137.053913
8534	Valence	60404.0	5923.0	90.873702
26156	Chambéry	57669.0	4753.0	86.607964
26365	Annecy	51423.0	5842.0	99.534074

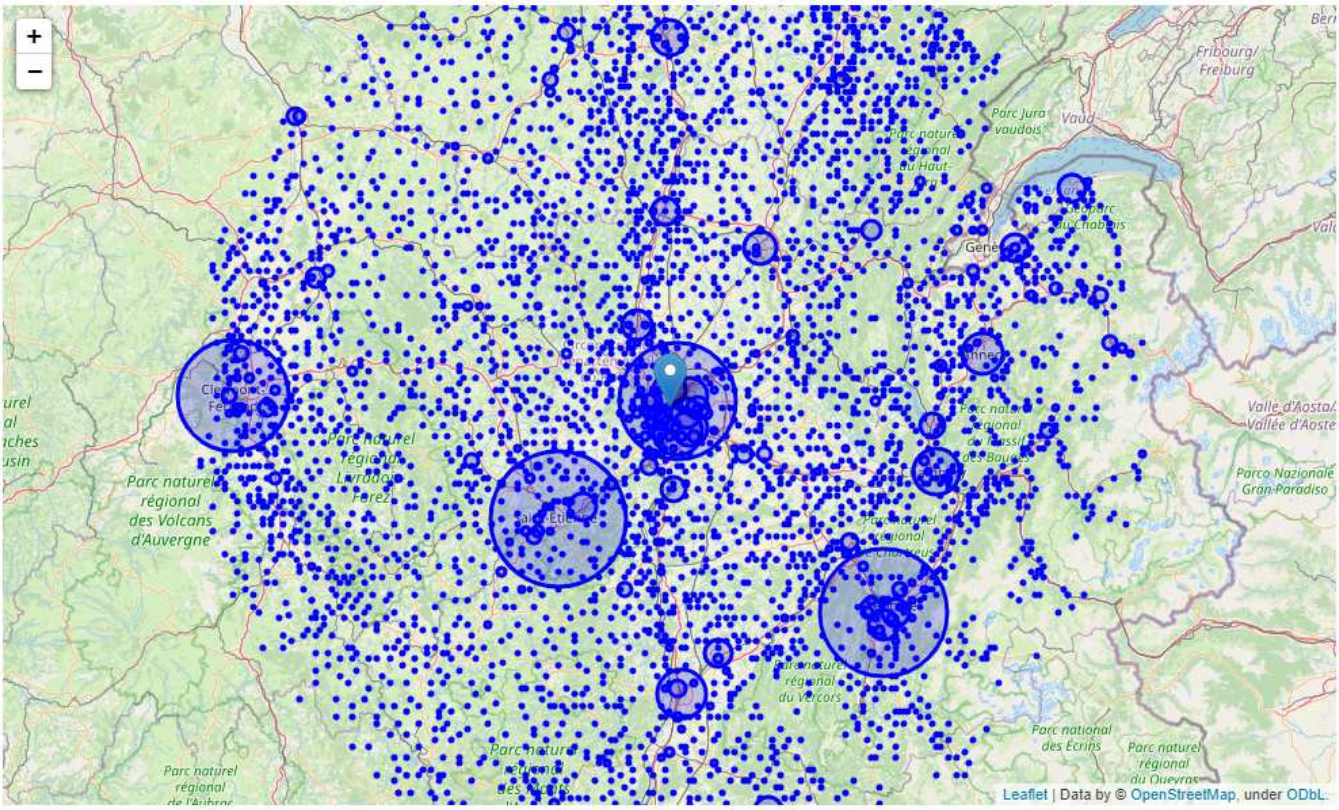
Réseau SNCF TGV, TER autour de Lyon



Réseau de Transport autour de Lyon (rail, route, air)



Répartition de la population :



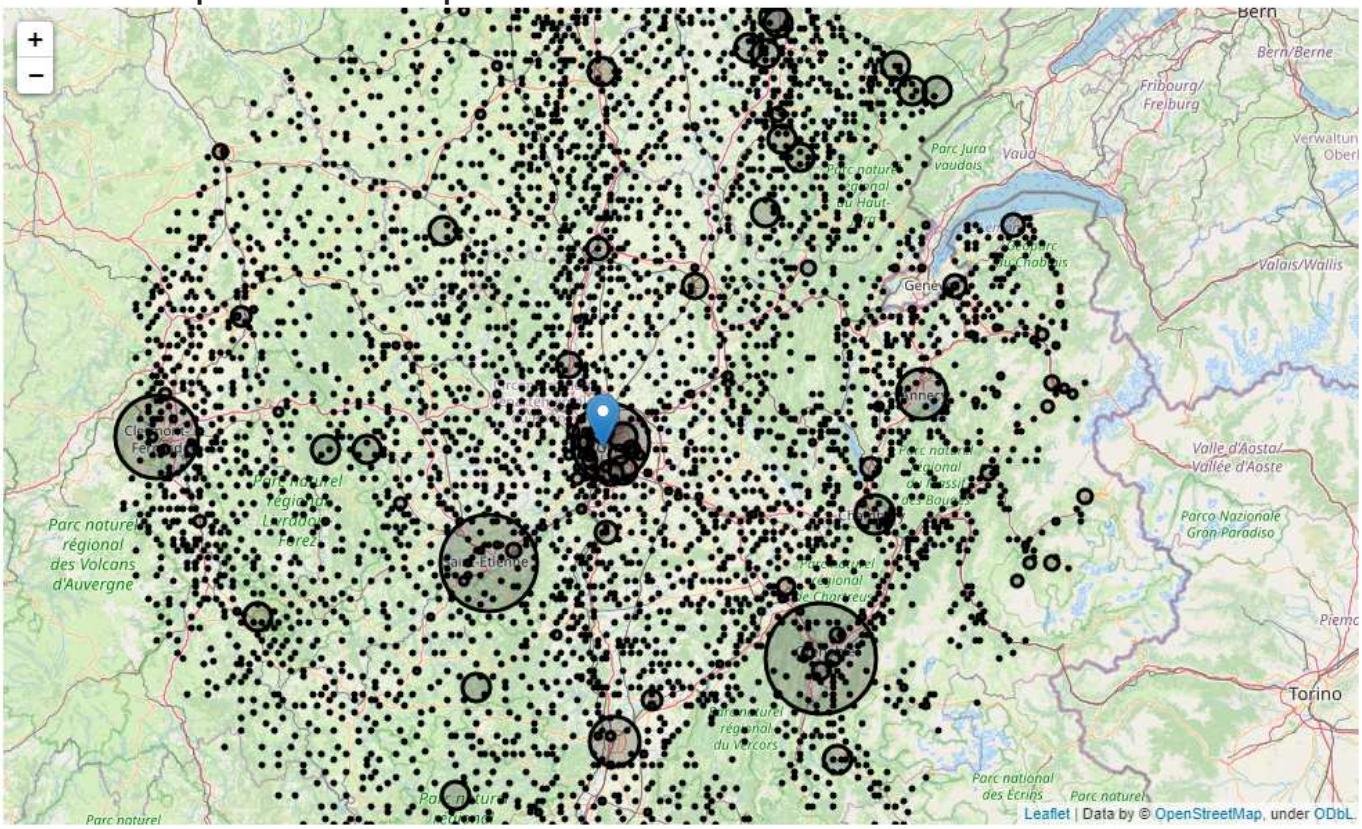
Répartition de la Population sur la zone étudiée :

Carte Open Street Map (Folium) - Population par commune dans un périmètre d'un rayon de 150 km autour de Lyon.

Note : La taille du point est proportionnelle au nombre d'habitants dans la commune.

Lyon et sa Métropole et Saint Etienne et Grenoble sont les 3 zones les plus peuplées de la zone définie. Clermont-Ferrand, à la limite de la zone étudiée a aussi une forte population. Puis on retrouve les villes avec une population entre 25 000 et 50 000 habitants, et situé à plus de 20 km de Lyon (Bourg-en-Bresse, Mâcon, Vienne.).

Répartition des Entreprises sur la zone étudiée :



Carte Open Street Map (Folium) - Nombre d'entreprises par commune dans un périmètre d'un rayon de 150 km autour de Lyon.

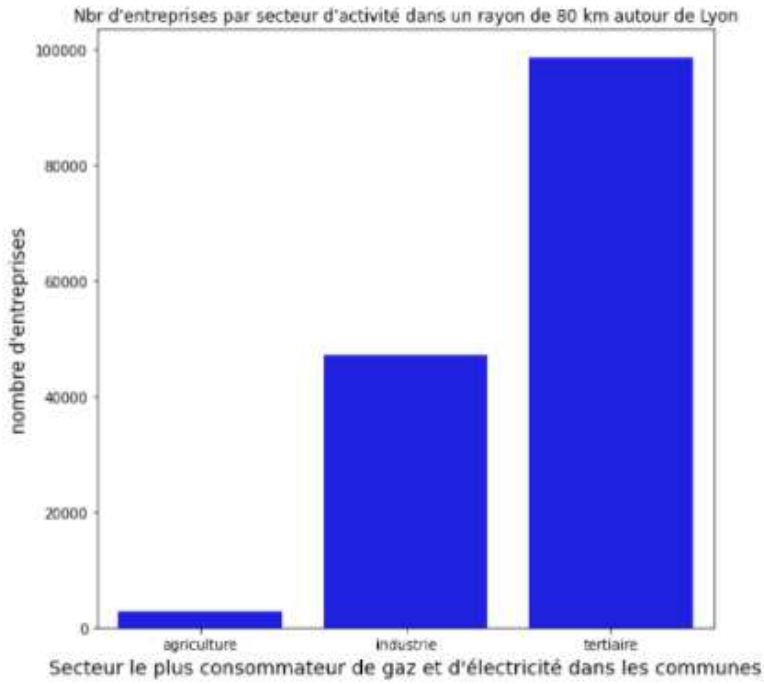
Note : la taille du point est proportionnelle au nombre d'entreprises sur la commune

On note une quasi-juxtaposition avec la carte précédente. Les villes avec une population supérieures à 25 000 habitants et situées à plus de 20 km de Lyon ont aussi une forte densité d'entreprises. La densité population semble être répartie de la même façon que la densité des entreprises sur le territoire.

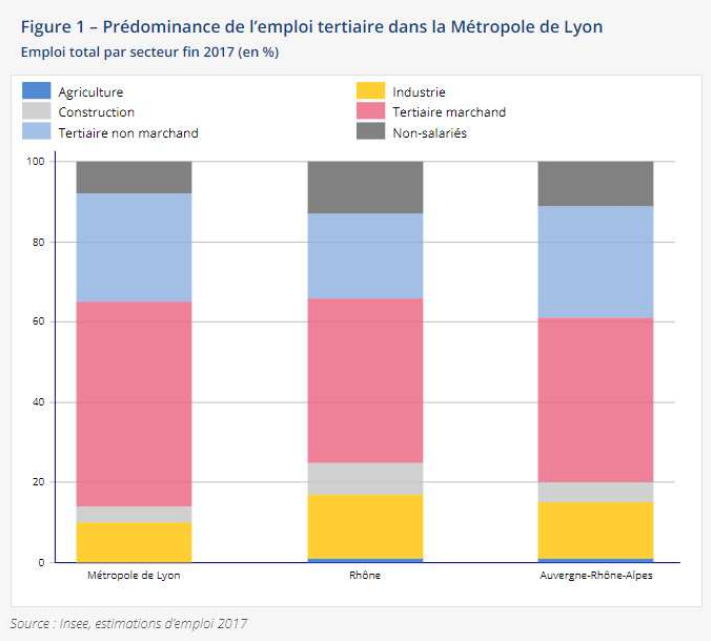
Mais comment sont répartis ces entreprises par secteur d'activité ?

4.2.1.2 Entreprises par secteur d'activité :

Comme dans la partie précédente, l'activité principale de la commune est déterminée en fonction du secteur d'activité consommant le plus de gaz et d'électricité

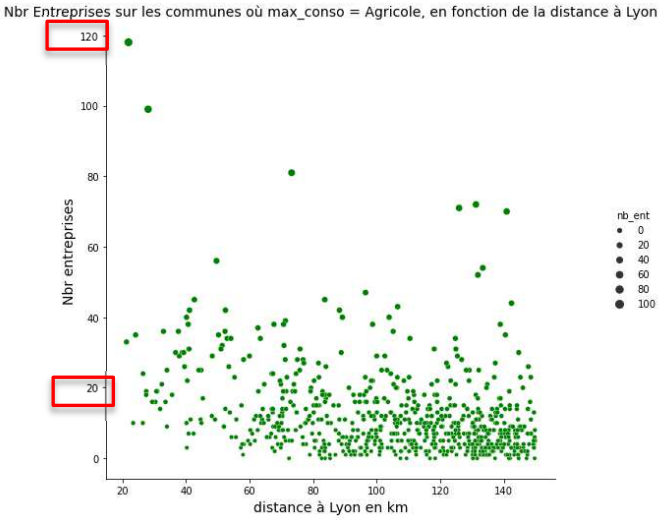


L'activité Tertiaire est la principale sur la zone étudiée.
Et comme l'indique le rapport Insee "**Département du Rhône : deux collectivités, un million d'emplois**" paru le 28/07/2020, il y a une prédominance du secteur de l'emploi Tertiaire dans la Métropole de Lyon, et en Rhône-Alpes. Il est important de comprendre comment sont répartis les entreprises de chaque secteur sur la zone étudiée.

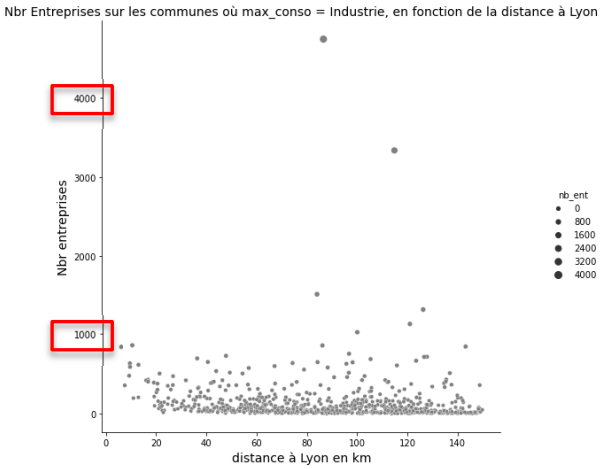


Les 3 graphiques montrent la répartition du nombre d'entreprises, pour chaque secteur d'activité, en fonction de la distance à Lyon.
On remarque que les échelles ne sont pas les mêmes, avec au maximum 120 entreprises pour le secteur Agricole, 900 pour le secteur Industriel, et 3500 pour le secteur Tertiaire.

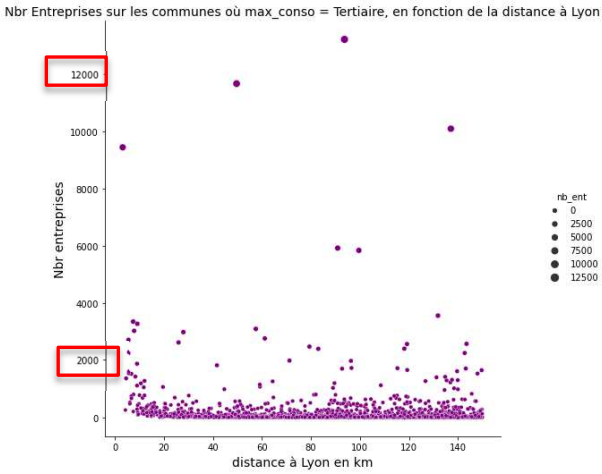
Communes dont le secteur d'activité avec la plus forte consommation de Gaz et d'électricité **= Agricole**



Communes dont le secteur d'activité avec la plus forte consommation de Gaz et d'électricité = **Industriel**



Communes dont le secteur d'activité avec la plus forte consommation de Gaz et d'électricité = **Tertiaire**



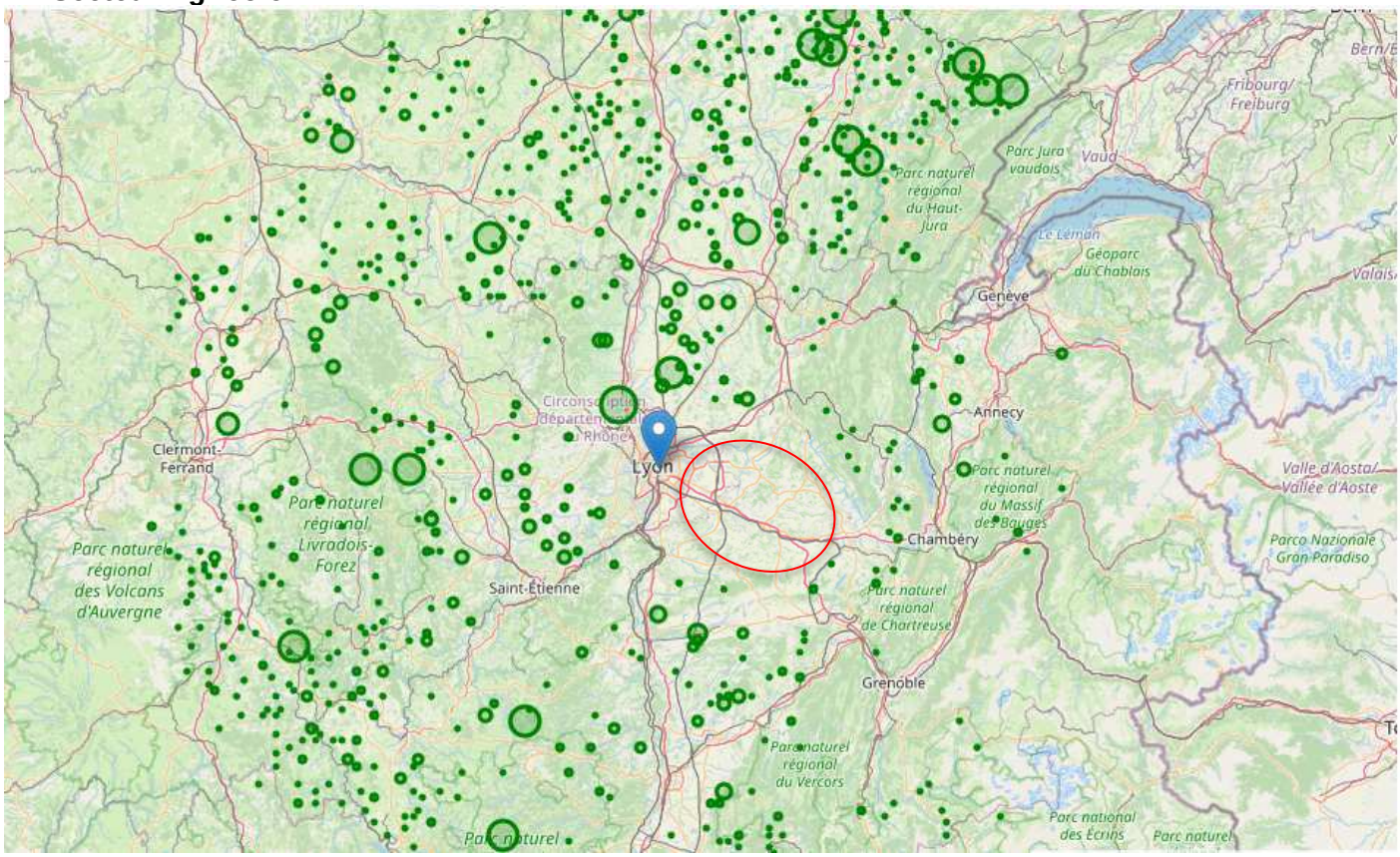
Les communes ayant une activité principalement agricole comptent, pour la plupart, entre 1 et 50 entreprises par communes et se situent à partir de 20 km de Lyon

Les communes ayant une activité principalement industrielle (d'après leur consommation en gaz/électricité) comptent entre 1 et 6000 entreprises. Chambéry est identifiée dans les communes dont l'activité Industrielle est celle consommant le plus de Gaz et d'Electricité, et compte près de 4700 entreprises. (Placoplâtre, Les Moulins de Savoie sont de gros sites industriels installés à Chambéry).

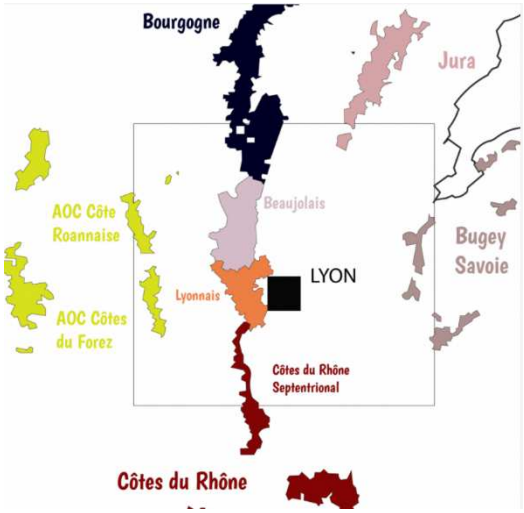
Les communes ayant une activité principalement Tertiaire (d'après leur consommation en gaz/électricité comptent entre 0 et 12 000 entreprises. Plus on s'éloigne de Lyon, plus le nombre de communes avec une nombre réduit d'entreprises (moins de 200) augmente.
On note un grand nombre de communes avec moins de 500 entreprises sur leur territoire, quel que soit la distance des communes à Lyon.
Grenoble, Saint-Etienne, Clermont-Ferrand comprennent plus de 10 000 entreprises. Elles sont également des préfectures, ce qui semble cohérant avec un secteur tertiaire prépondérant.

Ceci confirme donc la prédominance du secteur Tertiaire sur l'ensemble de la zone étudiée.

Secteur Agricole :



Carte Open Street Map (Folium) - Taille du point = Nombre d'entreprises dans les communes avec une activité majoritaire dans le secteur Agricole par commune dans un périmètre d'un rayon de 150 km autour de Lyon



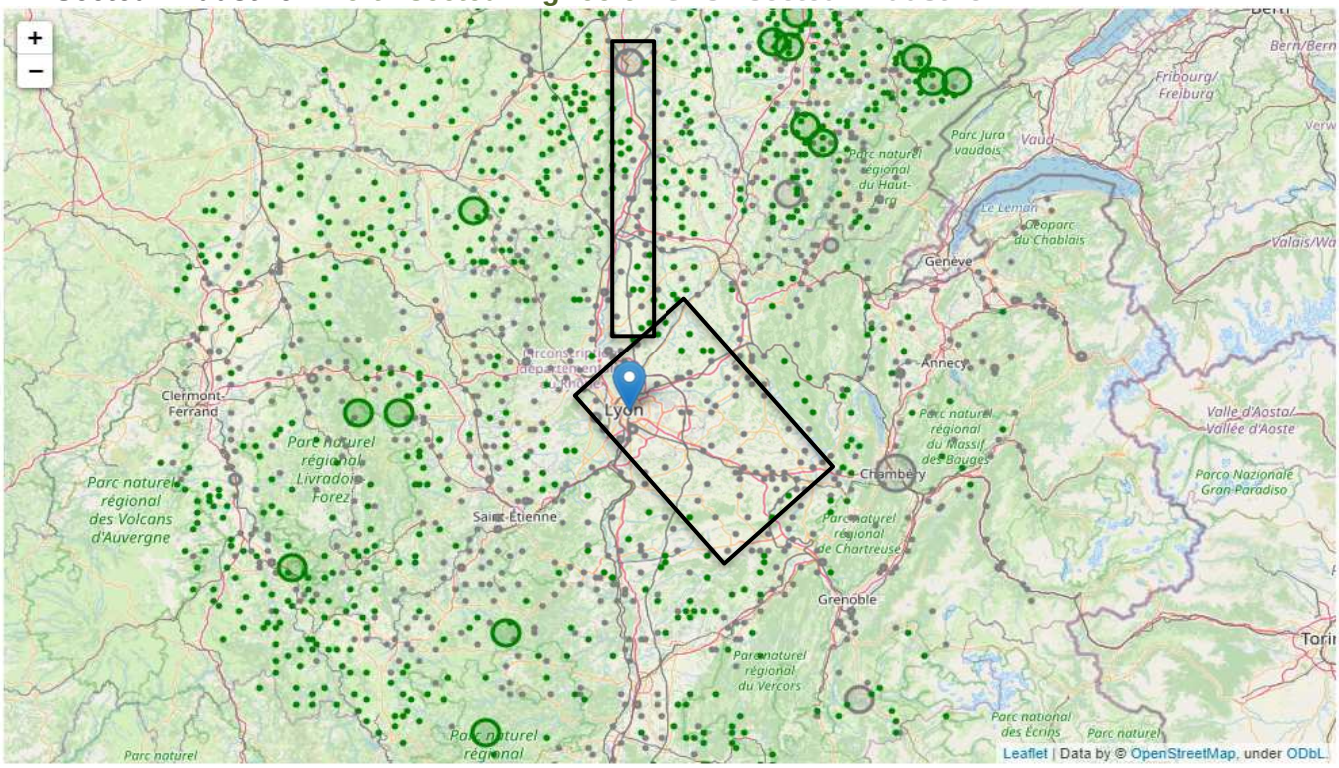
D'après La Chambre d'Agriculture Auvergne Rhône Alpes, une des activités agricoles principale est l'activité Viticole / Vinicole, traduit sur ces 2 cartes.

Au Nord Est du Parc Naturel Régional Livradois Forez, on a une concentration de domaine vinicole (appellation côte-de-forez) et d'activités laitières (zone laitière du Massif Central)

Au Nord du Parc Naturel Régional du Jura, on note un grand nombre d'entreprises, à priori dans le secteur d la production laitière, le Jura étant est un département agricole à vocation laitière dominante

On note également un “désert” agricole autour de Bourgoin-Jallieu : Pourquoi ? Il s’agit d’un pôle historique de l’industrie textile, et les activités Industrielles semblent rester prédominantes sur ce territoire. (voir secteur Industriel).

Secteur Industriel : vert : secteur Agricole Gris : secteur Industriel



Carte Open Street Map (Folium) - Taille du point = Nombre d’entreprises dans les communes avec une activité majoritaire dans le secteur Industriel par commune dans un périmètre d’un rayon de 150 km

On remarque au premier coup d’œil :

- Il ne semble pas y avoir d’activité Industrielle à Lyon même, mais à proximité (environ à 10 km de Lyon)
- Les Pôles industriels majeurs semblent se répartir le long des axes autoroutiers. 2 exemples

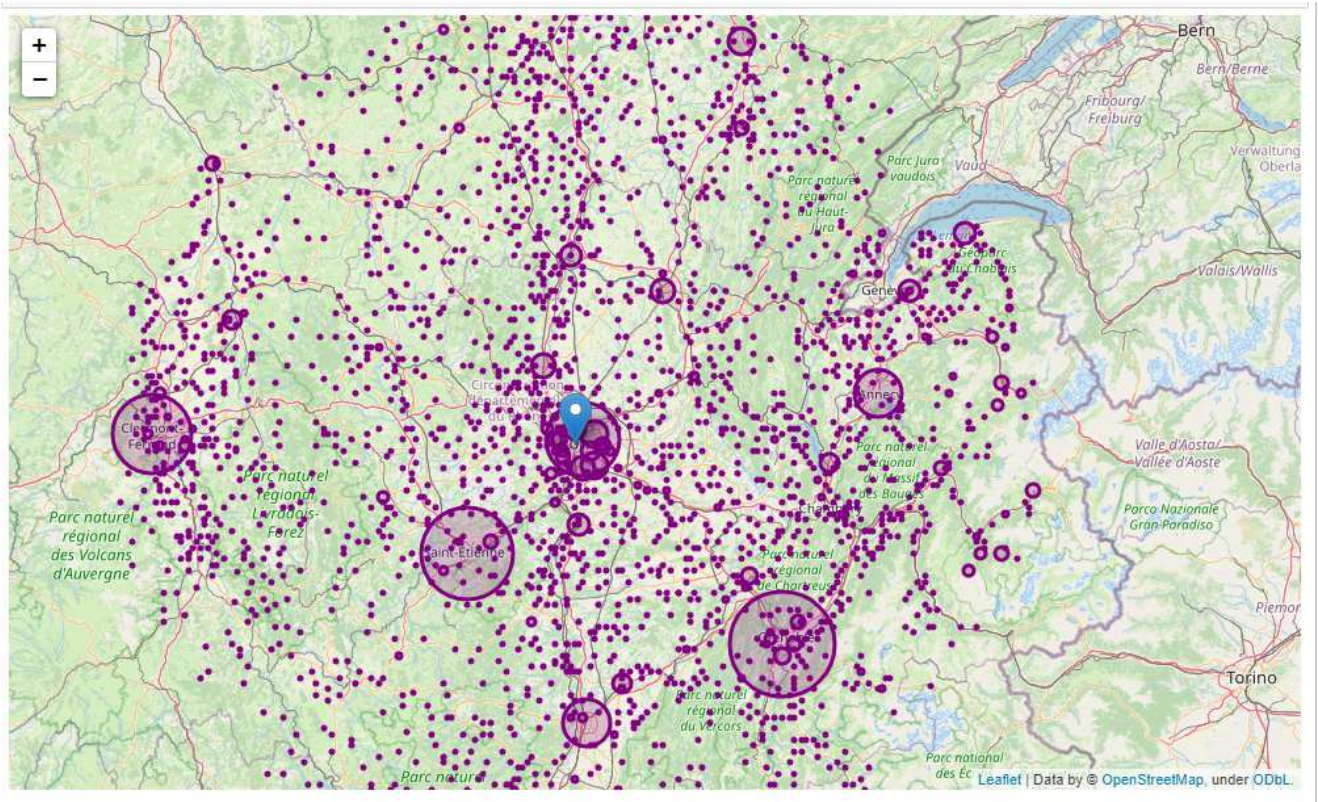
⇨ Au sud-est de Lyon, sur l’axe Lyon Chambéry Grenoble, le territoire autour de Bourgoin-Jallieu a une activité Industrielle quasi exclusive, et concentrée autour des villes principales (Axe Lyon - Chambéry, Ambérieux, Genas…)

Le rapport très complet sur l’urbanisme (lien ci-dessous) aide à mieux comprendre ce territoire. Au moment de la révolution Industrielle, haut-lieu de l’activité textile et industries associées (Mécanique, Bois, carton...) liées la “La Fabrique” Lyonnaise, l’axe Lyon - Chambéry - Grenoble se développe. Puis dans les années 70, la création de l’Aéroport Saint-Exupéry (à moins de 10 km à l’Ouest de Genas) y, et de l’autoroute 43 /48 permet un développement de ce territoire. Les activités industrielles sont d’ailleurs réparties autour de cet axe autoroutier. (ihttps://www.bourgoinjallieu.fr/images/2-Mairie/Urbanisme/Tome_1_diagnostic-avec_compression.pdf)

⇨ Au Nord de Lyon, il semble y avoir plutôt une activité Industrielle sur l'axe Lyon Mâcon, le long de la Saône, et/ou de l'autoroute A6 / A46.

Dans toute la région lyonnaise, l'industrie est un secteur économique majeur, du fait de l'Histoire de l'Industrie dans la Région et des innovation techniques. Les activités industrielles sont très diversifiées : de la soierie (secteur d'activité historique, en référence aux Canuts), la chimie, l'environnement, les transports, la mobilité, l'énergie ou encore la pharma-biotech.

Secteur Tertiaire :



Carte Open Street Map (Folium) - Taille du point = Nombre d'entreprises dans les communes avec une activité majoritaire dans le secteur Tertiaire. Par commune dans un périmètre d'un rayon de 150 km.

Les entreprises du secteur Tertiaire sont présentes sur l'ensemble du territoire étudié. Néanmoins, les communes comprenant le plus grand nombre d'entreprises du secteur Tertiaire sont :

- Concentrées dans Lyon et sa Métropole
- A Saint-Etienne, Grenoble, Clermont-Ferrand, Valence, soit les principales villes de la zone étudiées, qui sont également Préfectures de département.

Sur le reste du territoire, on note une multitude de communes avec une activité tertiaire principale, mais un nombre réduit d'entreprises.

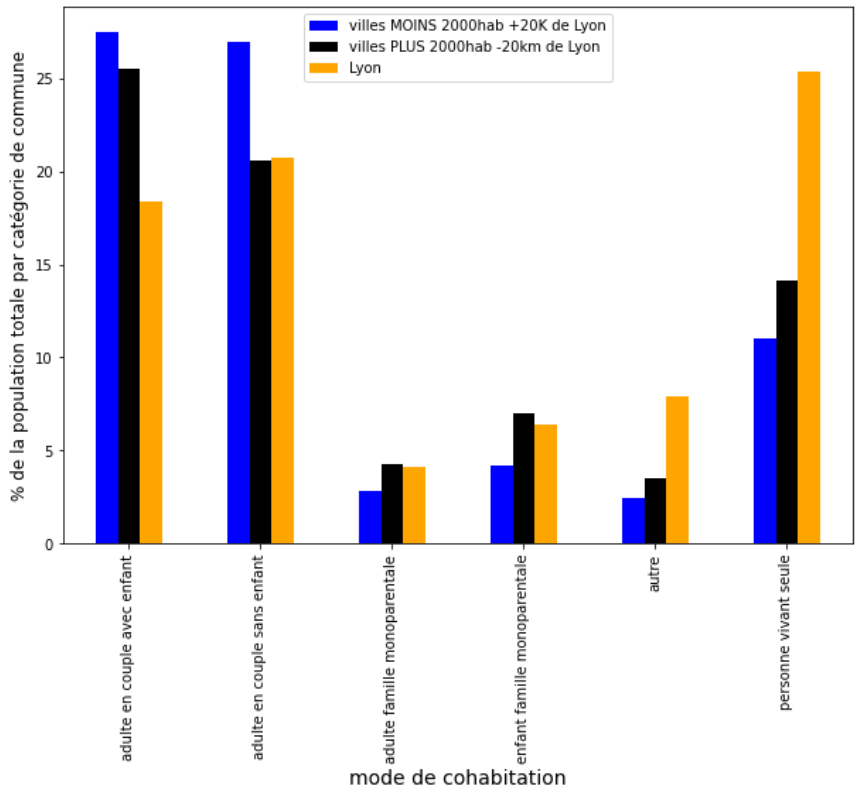
4.2.2 Population : répartition par Ages et mode de cohabitation

a) Mode de cohabitation – inégalités entre les grandes villes et les communes rurales

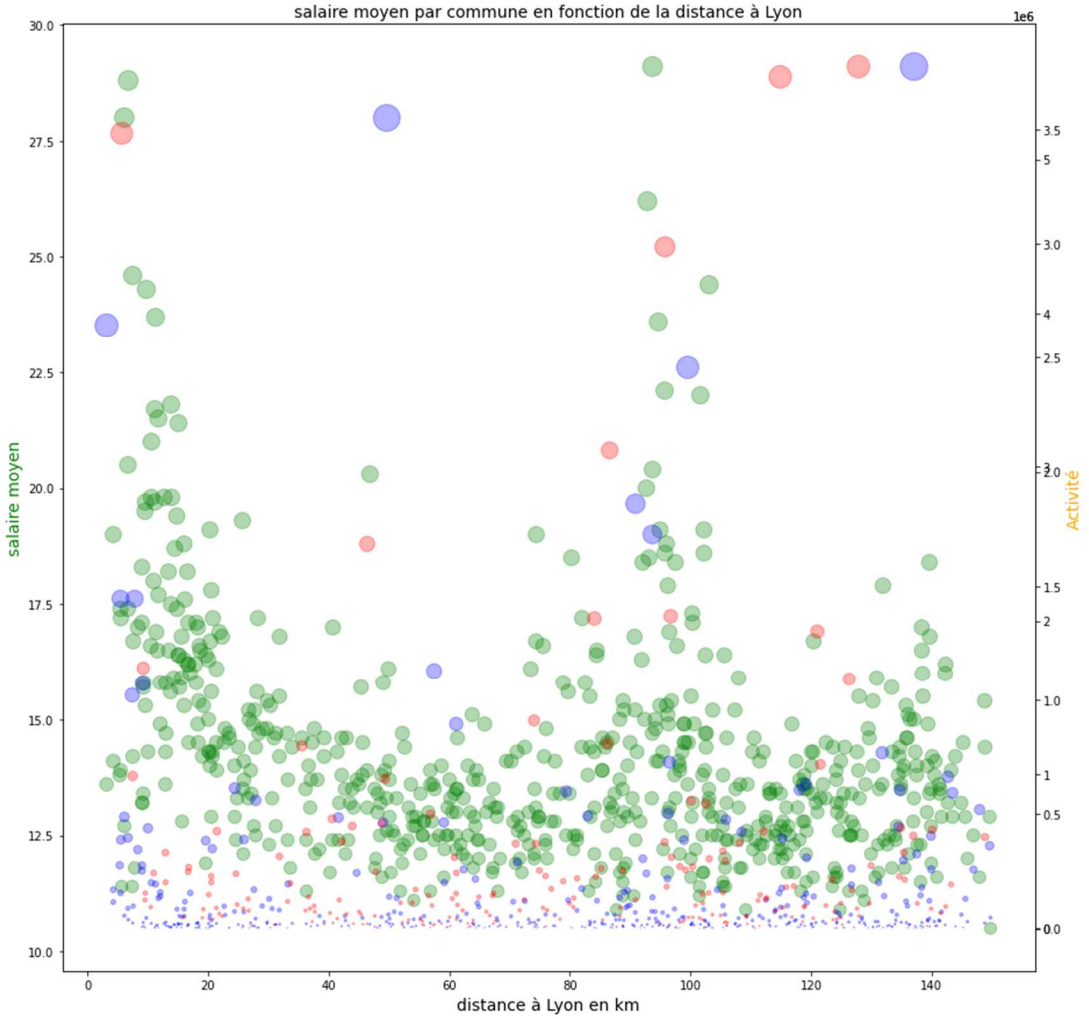
On cherche à comprendre si les modes de cohabitations principaux sont différents dans les communes dites « rurales », c'est-à-dire de moins de 2000 habitants et situées à plus de 20 km de Lyon, et les communes plus « urbaines » : plus de 2000 habitants et à moins de 20 km de Lyon, et à Lyon même.

On note une différence nette entre les modes de cohabitations principaux dans les zones « urbaines » et les zones « rurales ». Ces dernières comprennent plutôt des foyers en couple, avec et sans enfants, tandis que Lyon comporte majoritairement des personnes vivant seules (plus de 25% de la population Lyonnaise).
Il y a presque deux fois plus de familles monoparentales en zones « urbaines » (=Lyon et villes situées à moins de 20 km et de plus de 2000 habitants, barres noire et orange cumulées) qu'en zones « rurales » (villes de moins de 2000 habitants et à plus de 20 km de Lyon).

Et à Lyon même, il y a près de 3 fois plus de personnes vivant seules que dans les communes de moins de 2000 habitants et situées à plus de 20 kilomètres de Lyon.



4.2.3 Niveau de Salaire en fonction de la distance à Lyon.

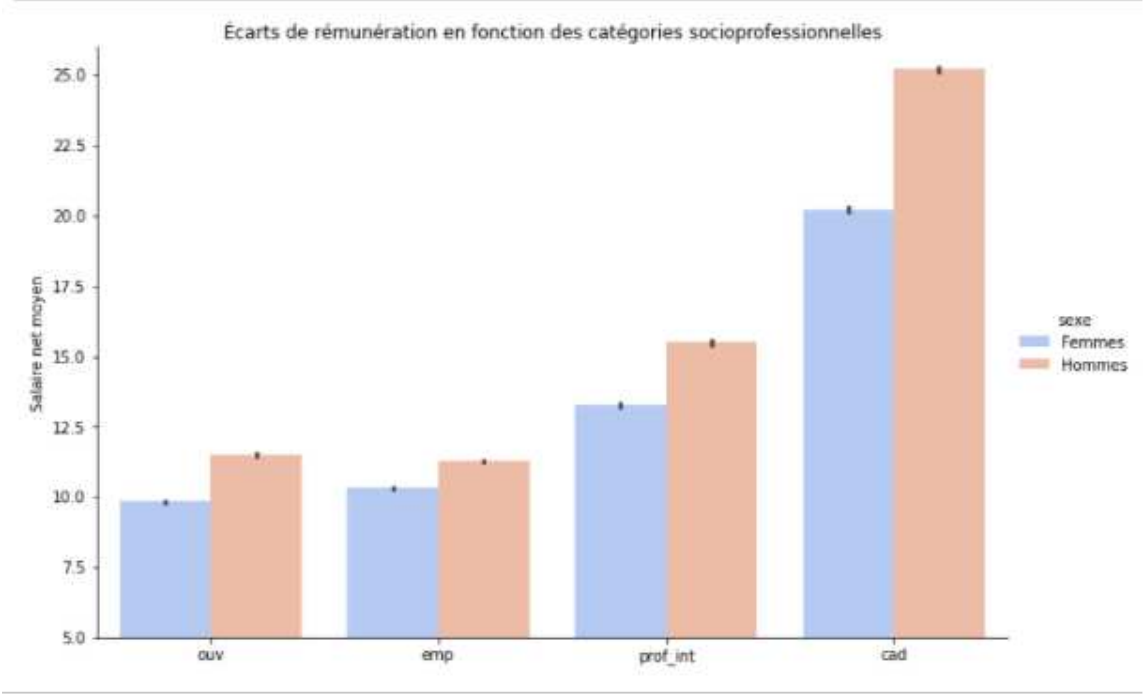


Les points en vert sont de même taille et indiquent le niveau du salaire moyen représenté par l'axe des ordonnées à droite.
Les points en bleu indiquent l'activité tertiaire et ceux en rouge indiquent une activité industrielle. La taille en bleu et en rouge varie en fonction du niveau de la consommation représenté par l'axe des ordonnées à gauche.

Comme pour lors de la visualisation de la population et du nombres d'entreprises par ville, en fonction de la distance à Lyon, on note des salaires moyens plus élevés à Lyon et à proximité, puis une diminution du salaire moyen (hors à Saint-Etienne) , et enfin, à proximité de Grenoble (93 km de Lyon), puis de Clermont-Ferrand (137 km de Lyon) des salaires moyens à nouveaux plus élevés, c'est-à-dire supérieurs à 17,5€ / heure. On confirme donc que la distance des communes à un Métropole a un impact sur la population des communes, le nombres d'entreprises présentent et également le niveau de salaire moyen, jusqu'à « se rapprocher » d'une autre Métropole.

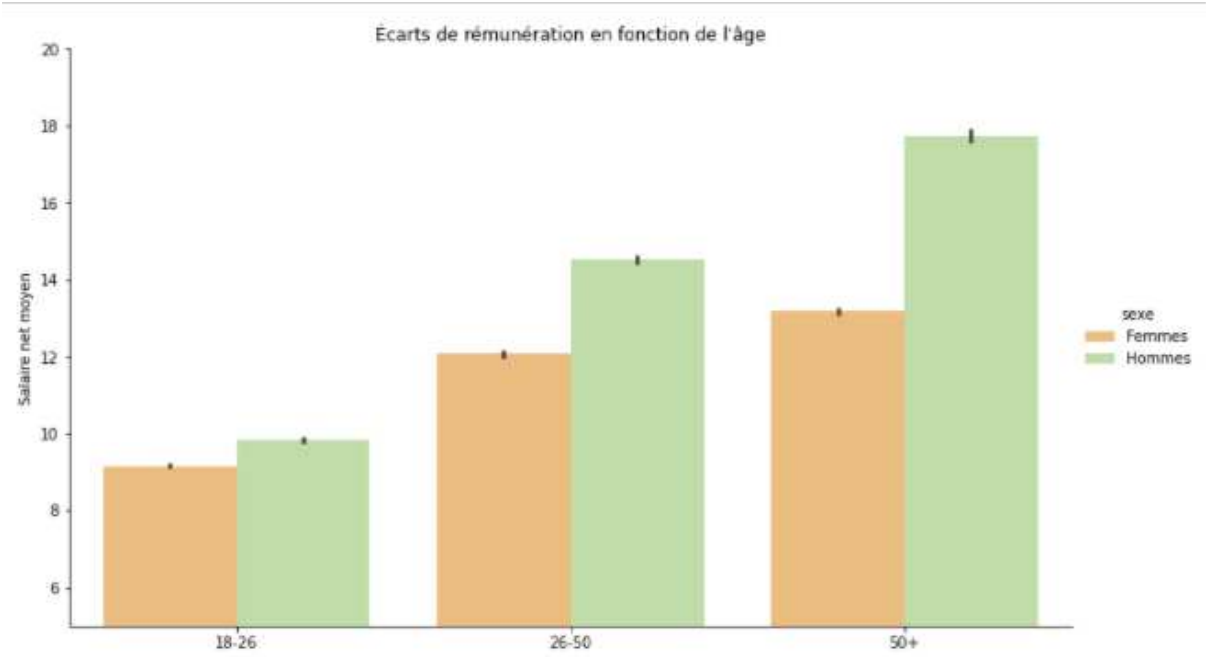
4.3 Ecarts de rémunération Femmes / Hommes

4.3.1 Ecart dans les catégories socioprofessionnelles



Le jeu de données sur le salaire moyen illustre bien les inégalités professionnelles entre les femmes et les hommes à travers le graphique ci-dessous.
L'écart semble encore plus élevé chez les cadres.

4.3.2 Ecart par à l'âge



L'écart semble aussi présent dans tous les groupe d'âge mais il est moins net par rapport à la population la plus jeune.
Il est probablement que cet écart augmente avec l'âge mais cela nécessite d'autres analyses et d'autres données pour vérifier cette hypothèse.

5 Modélisation

Dans le Chapitre 4.2, nous avons vu que les communes dans un périmètre de 80 km autour de Lyon, en fonction de leur distance à Lyon, peuvent avoir des « profils » assez différents.
Par exemple, plus on s'éloigne de Lyon et moins la population et le nombre d'entreprises dans ces communes est important.
Et quand le périmètre est de 140 km, il est possible de visualiser que la proximité d'une autre grande ville (Grenoble par exemple), on note à nouveau des communes avec un plus grand nombre d'entreprises et des population plus importantes.

Aussi, il serait intéressant de voir les résultats d'un calcul de Clusters dans une zone géographique de 140 km autour de Lyon.
L'exercice de calcul de Cluster sera fait sur le jeu de données fusionnées, sauf les données de Salaire, soit le fichier master.csv.

Les communes dans un périmètre de 140 km autour de Lyon seront sélectionnées et Lyon exclus (les valeurs des variables sur Lyon sont très élevées par rapport aux valeurs des variables des communes en périphérie de Lyon).

5.1 Calcul des Clusters sur les communes situées dans un périmètre de 150 km autour de Lyon – sans la variable Salaire

5.1.1 Sur le jeu de données Master : Pré-traitement

5.1.1.1 Chargement des données

Les étapes sont :

- Chargement du jeu de données
- Création d'un filtre « universel » pour sélectionner la ville cible et le périmètre cible :
  ```
  commune = 'Lyon'   # définit Lyon comme centre de cercle
  distance = 150      # distance  max  à Lyon en km
  ```

- Mise en œuvre de la fonction « sperical_dist » permettant de calculer la distance entre 2 communes et définie dans un Notebook appelé Utils :

- **Le Dataframe df_lyon comporte 4069 communes**, et bien sûr les 65 variables du jeu de données master. L'échantillon parait assez important pour les prochaines étapes, en vue de la mise en œuvre d'une analyse des composantes principales et d'un modèle de Clustering KMeans.

5.1.1.2 Pré-traitement

Contenu de df_lyon : cf notebook 11bis.
Le jeu de données comporte 75 variables décrivant les 4069 communes du périmètre choisi.

Entre autres, le jeu de données comprend le nombre total d'entreprises par communes et leur nombre en fonction de la taille de l'entreprise (nombre de salariés).
Il en est de même pour la Population Totale et la population par tranche d'âge, par genre, par mode de cohabitation.
Et enfin pour la Consommation totale (Gaz / Electricité) et les consommations des Secteur d'activité Tertiaire, Industriel et Agricole.

Plusieurs tentatives de mises en œuvre du modèle PCA puis KMeans ont finalement montrées que les modèles réagissent mieux quand les fréquences de chaque catégorie sont appliquées au jeu de données plutôt que les chiffres bruts :

Pour les entreprises, par exemple, **plutôt que de considérer le nombre d'entreprises par taille, on va considérer la part représentée par chaque catégorie d'entreprise.**
Et on fait de même pour la population par tranche d'âges, par genre, par mode de cohabitation.

Et enfin, considérer la part des consommations des secteurs d'activité par rapport à la consommation totale (Gaz / Electricité) de la commune.

```
# 1 - pour entreprises et nbr d'entreprises par taille (nbr de salariés)
for col in df_lyon.columns:
    if 'nb_ent_' in col:  # si nb_ent qui est dans le nom de la colonne, alors on divise la valeur de
nb_ent_0 par nb_ent
        df_lyon[col] = df_lyon[col] / df_lyon['nb_ent']

# 2 - pour population et population par tranche d'âge
for col in df_lyon.columns:
    if 'pop_age_' in col:
        df_lyon[col] = df_lyon[col] / df_lyon['pop_total']

# 3  - pour population et genre
for col in df_lyon.columns:
    if 'pop_sexe_' in col:
        df_lyon[col] = df_lyon[col] / df_lyon['pop_total']

# 4 - pour population et mode de cohabitation
for col in df_lyon.columns:
    if 'pop_mode_cohabitation_' in col:
        df_lyon[col] = df_lyon[col] / df_lyon['pop_total']
# 5 - pour conso
for col in df_lyon.columns:
    if 'consommation_libelle_grand_secteur' in col:
        df_lyon[col] = df_lyon[col] / df_lyon['consommation_totale']
```

- **Préparer un Dataframe ne contenant que des variables numériques.**

  En vue de la mise en œuvre du modèle PCA, il faut produire un Dataframe ne contenant que les données numériques.
  Le calcul précédent a produit des lignes NaN (valeur divisée par Zéro). On supprime des lignes NaN.

  On identifie donc les variables par datatype :

  Et on enlève toutes les variables de type « objet ».
  Les variables latitude, longitude, code_departement, distance_to_Lyon sont des variables numériques mais non explicatives pour les communes.
  On les enlève également.

  On va aussi éliminer les 'latitude', 'longitude', 'code_departement', 'distance_to_Lyon'car ces variables
  Elles ne sont pas utiles pour la PCA , ce ne sont pas des variables explicatives.
  On enlève aussi toutes les variables à propos du nombre de points,

  Enfin, on élimine également les variables relatives aux relevés de consommation (nombre de points ; consommation Gaz ou Electricité) pour ne garder que les consommations sur les secteurs d'activité tertiaire, industriel et agricole.

On obtient le df_lyon_num, qui comprend 41 variables, 4047 lignes.

Les échelles de valeur des variables sont très hétérogènes. On va donc normaliser.

➔ Normalisation avec MinMaxScaler

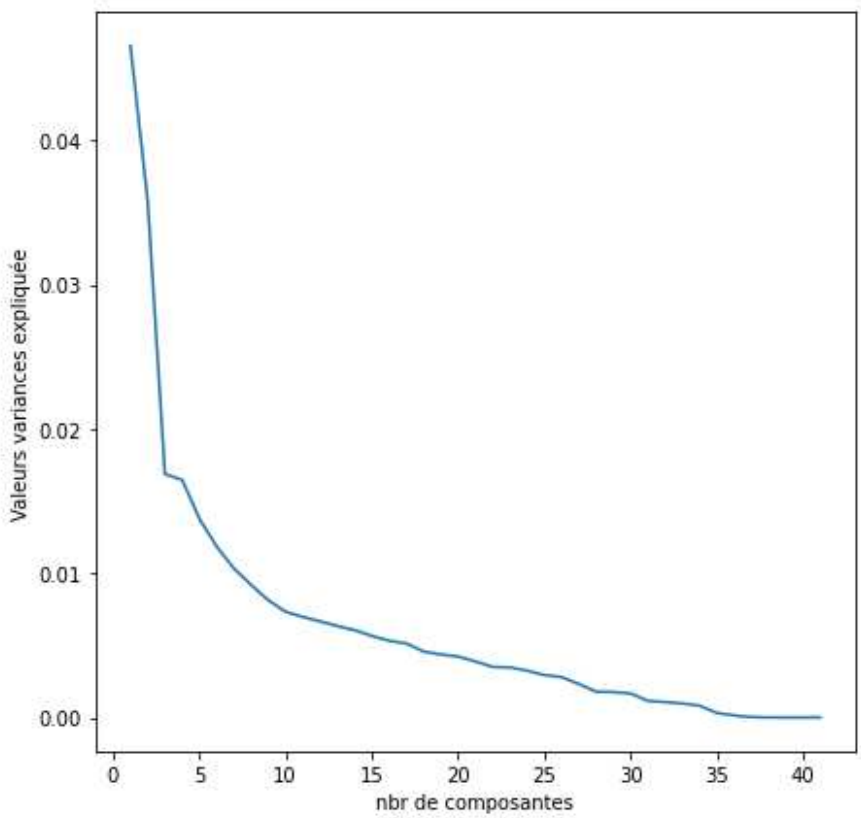
```
# normalisation de df_lyon_num avec MinMaxScaler:
# création du scaler

scaler = MinMaxScaler()

# normalisation de df_lyon_num :
df_lyon_num_scaled = scaler.fit_transform(df_lyon_num)
```

5.1.2 Mise en œuvre de l'Analyse des Composantes Principales (PCA)

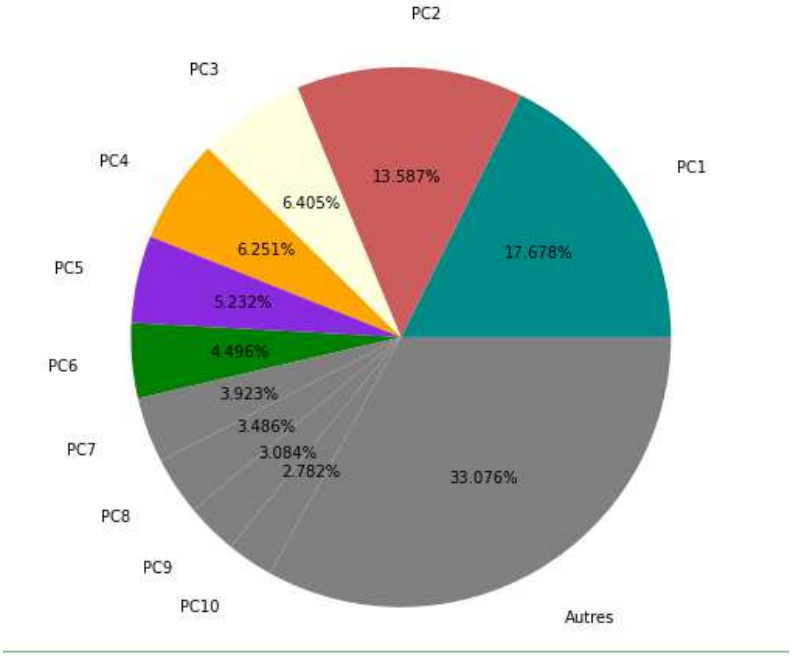
➔ Instanciation de PCA et calcul de coordonnées

```
# Instanciation d'une PCA

pca = PCA()

# l'intance pca est appliquée aux données avec la méthode fit_transform() (applique la réduction de dimension sur le np array)

coord = pca.fit_transform(df_lyon_num_scaled)
```

➔ Variance expliquée et nombre de composantes principales



On a clairement un « coude » à avec 2 composantes principales.
Qu'en est-il des ratios des variances expliquées ?

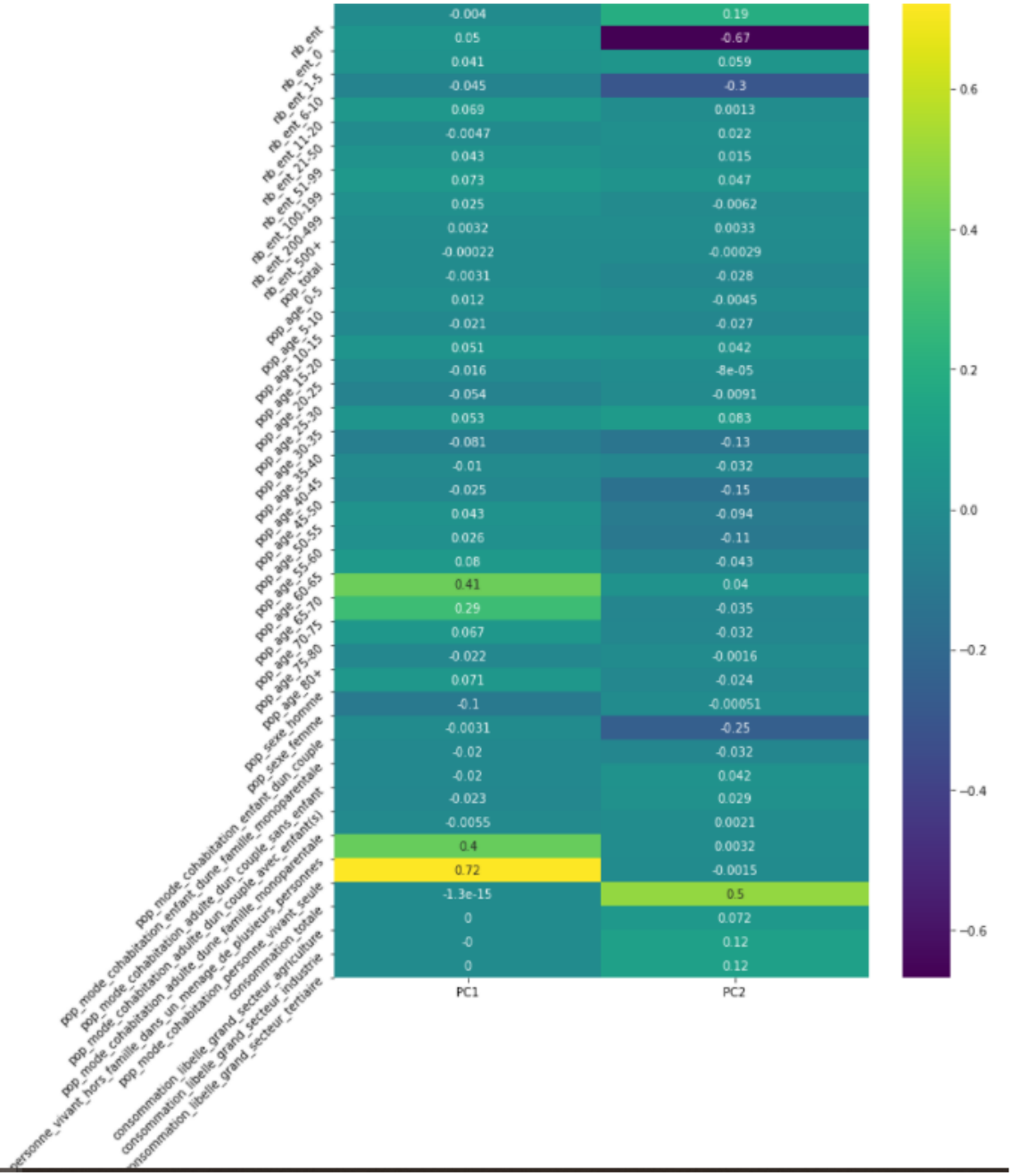


A partir de 5 composantes principales, on est à 50 % des ratios de variances expliquées cumulées.
Avec 2 composantes principales on est à 30% de variance expliquée cumulée,
Mais les ratios de variance expliqués des 2 premières Composantes Principales est égale à plus du double du ratio des autres composantes. Aussi on choisira de conserver 2 Composantes Principales.

➔ Corrélation des 2 Composantes principales avec les 41 variables

- Afficher les coefficients de corrélations de chaque variable
- Avec les deux premières CP (les 2 premiers axes)
- Grâce à l'attribut components_ de PCA

```
Comp_PCA = pd.DataFrame({'PC1': pca.components_[:, 0], 'PC2': pca.components_[:, 1]})

plt.figure(figsize=(10, 15))

sns.heatmap(Comp_PCA, annot=True, cmap='viridis')
plt.yticks(np.arange(1,42), df_lyon_num.columns, rotation = 45)
plt.show()
```



La variable « personnes vivants seules » est la plus corrélées avec la PC1, et la variable part des entreprises de 0 salariés est la plus corrélés (inversement) avec la PC2.

5.1.3 Mise en œuvre du Modèle de Clustering KMeans

➔ Création du Dataframe avec les 2 composantes principales calculées précédemment

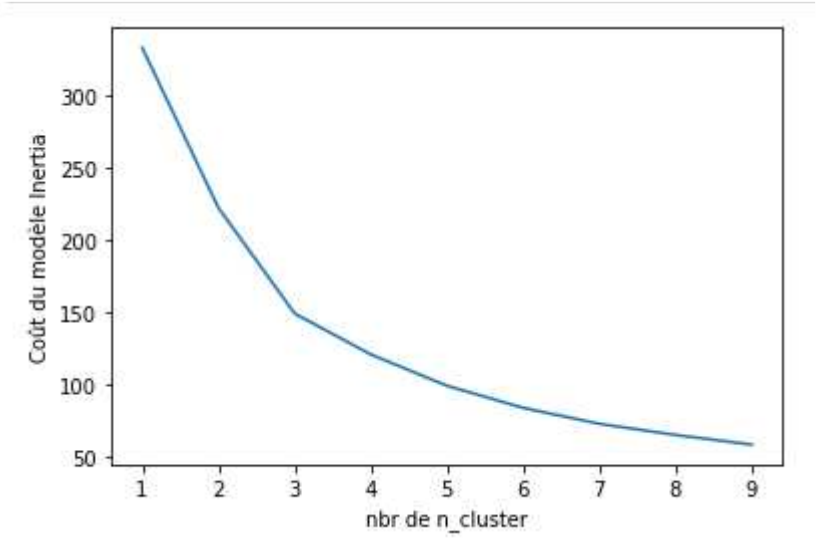
```
# mise en oeuvre de K-means à partir des 2 composantes principale , Comp_PCA
pca_lyon_kmeans_2PC = pd.DataFrame({'Axe 1': coord[:,0], 'Axe 2' : coord[:,1]})
print(pca_lyon_kmeans_2PC.shape)
```

➔ Instanciation d'un modèle de KMeans et méthode dite du coude pour évaluer le nombre de Clusters à considérer en hyperparamètre

```
# instanciation de kmeans - essai avec 4 clusters
kmeans_2PC = KMeans(n_clusters = 4)
```

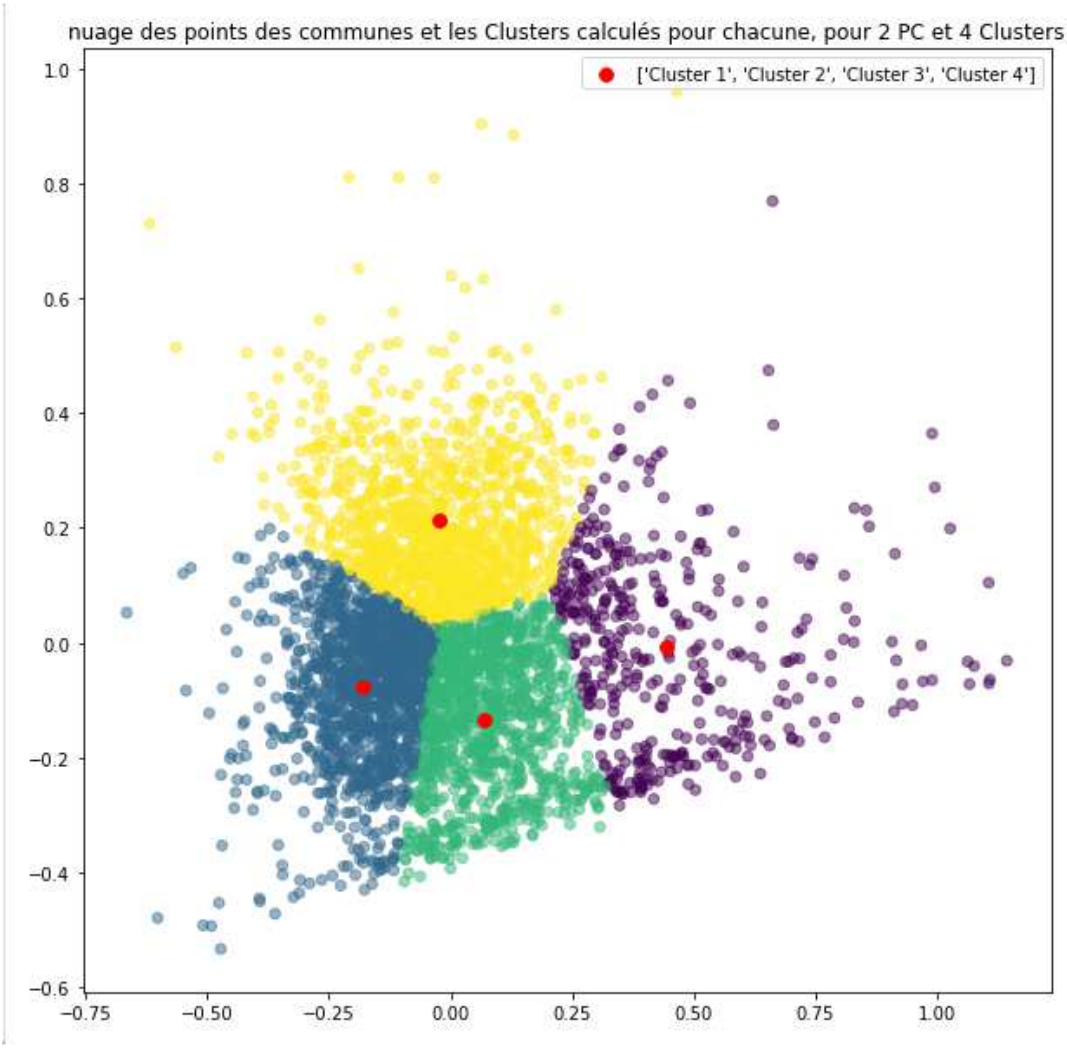
```
# entrainement sur Comp_PCA
kmeans_2PC.fit(pca_lyon_kmeans_2PC)

# Prédictions
y_kmeans_2PC = kmeans_2PC.predict(pca_lyon_kmeans_2PC)
print(y_kmeans_2PC)

# coordonnées des centres des clusters
kmeans_2PC.cluster_centers_

x_cluster_centers = kmeans_2PC.cluster_centers_[:,0]
y_cluster_centers = kmeans_2PC.cluster_centers_[:,1]

print(x_cluster_centers)
print(y_cluster_centers)

print("la valeur de l'inertie por le modèle est : ", kmeans_2PC.inertia_)
```

Inertia et méthode du coude :



Le nombres de n_clusters à régler en hyperparamètre peut être compris entre 2 et 4 d'après le graphique

➔ Tentative avec 4 clusters :



Il semble que 4 Clusters soit un bon choix. Bien que les Clusters « Jaune » et « Violet » montrent une moins bonne concentration des points autour du Centre de Cluster.

➔ **Création d'un code « universel » afin de tester plusieurs hypothèses**.

- Combien de Composantes principales faut-il prendre en compte ?
- A quelle valeur régler l'hyperparamètre n_clusters du modèle KMeans ?

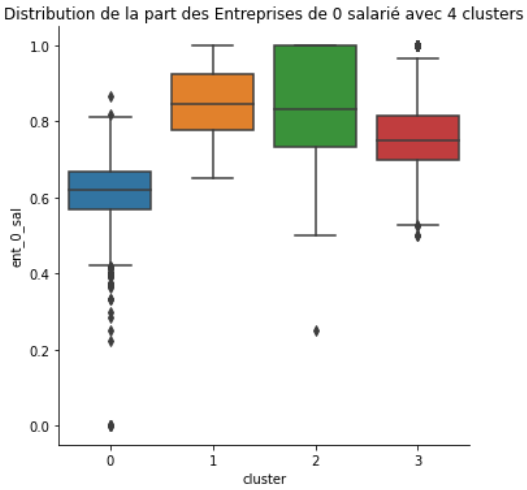
```
Création des variables :
n_variable pour fixer le nombre de Composantes principlaes à considérer
n_cluster pour fixer la valeur de l'hyper paramètre :

# Universalisation de la mise en oeuvre du modèle de KMeans
pca_lyon_kmeans = pd.DataFrame({"Axe " + str(i + 1): coord[:,i] for i in range(n_variable)})

# instanciation de kmeans - essai avec 3 clusters
# kmeans = KMeans(n_clusters = 3)
kmeans = KMeans(n_clusters = n_cluster)

# entrainement sur Comp_PCA
kmeans.fit(pca_lyon_kmeans)

# Prédictions
y_kmeans = kmeans.predict(pca_lyon_kmeans)
print(len(y_kmeans))

# coordonnées des centres des clusters
kmeans.cluster_centers_

x_cluster_centers = kmeans.cluster_centers_[:,0]
y_cluster_centers = kmeans.cluster_centers_[:,1]
```

Après plusieurs essais, il semble que 2 composantes principales et un hyperparamètre n_cluster fixé à 4 répondent assez bien :

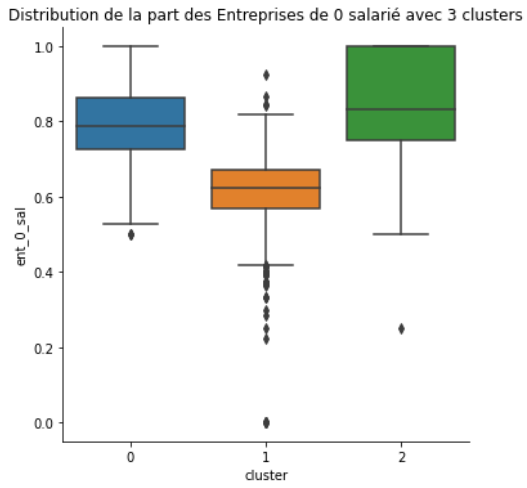
5.1.4 Comparaison de la distribution des 2 variables les plus corrélées les 2 composantes principales dans 2 cas de figure

1 – Distribution de la part des Entreprises de 0 salarié

a) Nombre de composantes = 2 ; nombre de clusters = 4



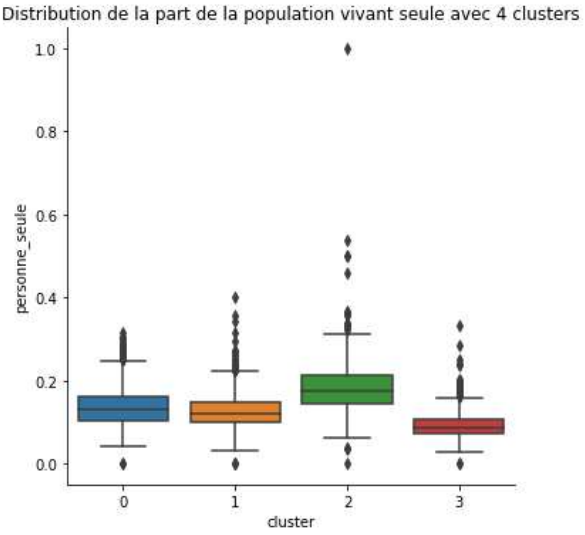
b) Nombre de composantes = 10, nombre de clusters = 3



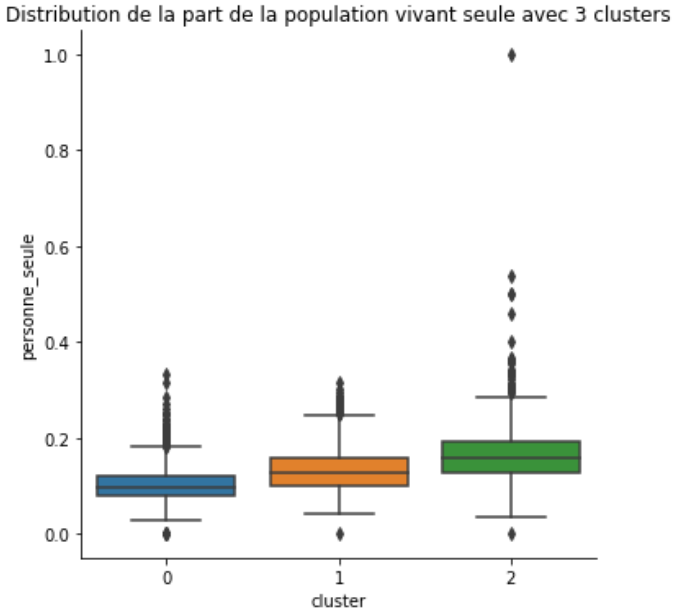
Dans les deux cas on note le même nombre de points « Outliers » pour le cluster 0 (a) et le cluster 1 (b)

2 – Distribution de la part de la part de la population vivant seule

a) Nombre de composantes = 2 ; nombre de clusters = 4



b) Nombre de composantes = 10, nombre de clusters = 3



De même, pour la distribution de cette variable, un nombre important de points « outliers » sont présents dans les 2 cas.

De plus avec seulement 3 clusters, on perd de la finesse dans la segmentation des communes. En effet, avec 4 clusters on identifie bien une « catégorie » de villes où la part de personnes vivant seules est inférieure 10% cluster 3 cas a).

Aussi, la conformation « 2 composantes principales et 4 clusters » est conservée.

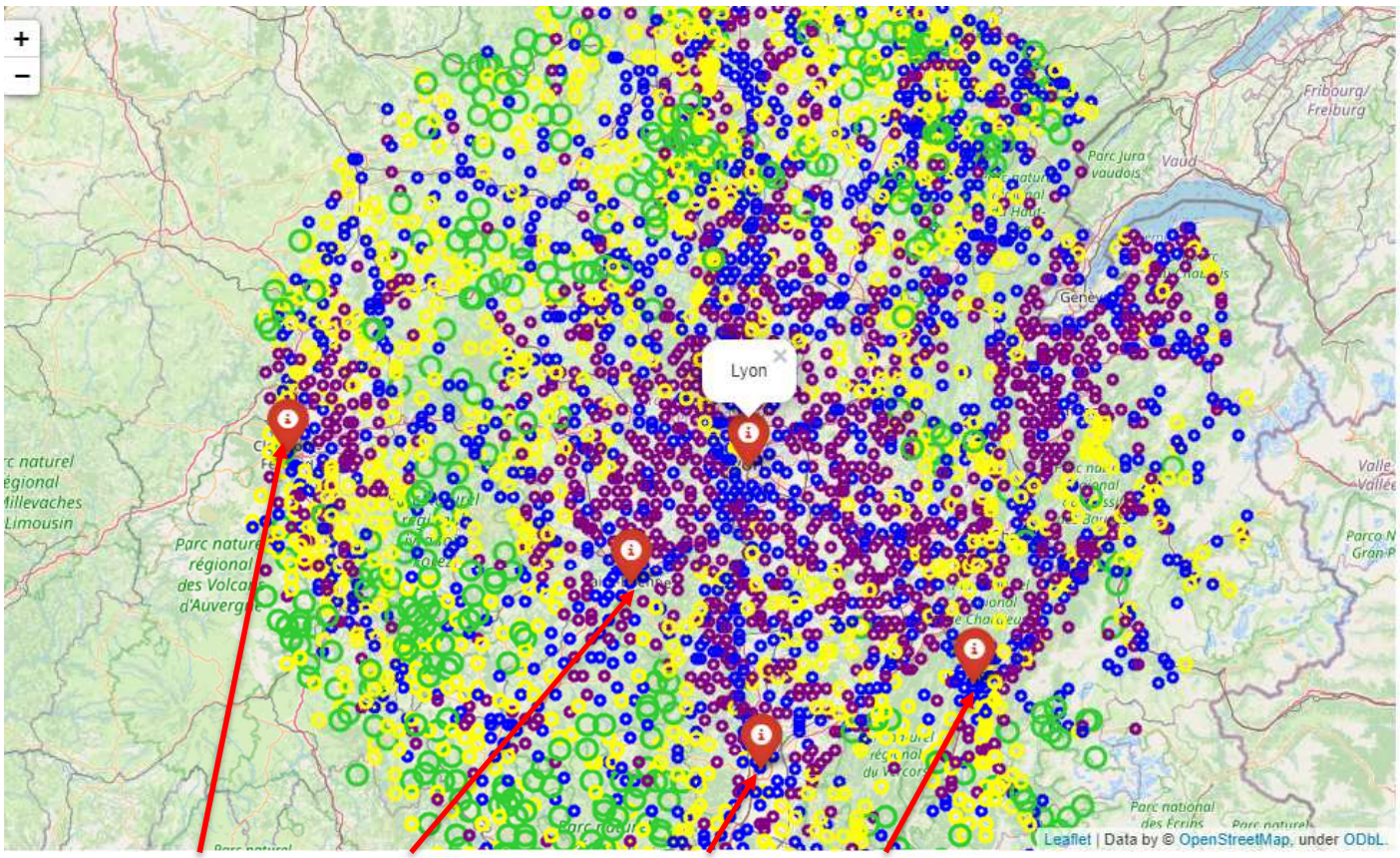
5.1.5 Comment les villes, réparties en 4 Clusters sont-elles situées sur une carte ?

Visualise-t-on des Clusters différents en fonction de la distance de chaque commune par rapport à une Métropole comme Lyon, ou une commune de grande taille come Saint-Etienne ou encore Grenoble ?

On utilise le package folium de python.

Visualisation des villes segmentées par Clusters, dans un périmètre de 140 km autour de Lyon :

Pour lire plus facilement la carte :
Légende :
Cluster 0 **Cluster 2**
Cluster 1 **Cluster 3**



Clermont-Ferrand Saint-Etienne Valence Grenoble

Statistiques pour le Cluster 0 : Grandes Villes / max. 166 137 habitant (Saint-Etienne)

	cluster	ent_0_sal	nb_ent	personne_seule	population_total	pop_enfants_0_20	pop_adultes_20_40	pop_adultes_40_60	pop_adultes_60_80plus
count	1202.0	1202.000000	1202.000000	1202.000000	1202.000000	1202.000000	1202.000000	1202.000000	1202.000000
mean	1.0	0.609989	286.452579	0.136352	3820.738769	0.239925	0.206690	0.289057	0.264328
std	0.0	0.095088	803.607250	0.047184	10594.755328	0.039617	0.050939	0.052038	0.065752
min	1.0	0.000000	1.000000	0.000000	21.000000	0.000000	0.000000	0.058824	0.050633
25%	1.0	0.568142	28.000000	0.103101	489.000000	0.216272	0.176296	0.259223	0.220452
50%	1.0	0.621434	95.000000	0.129673	1276.500000	0.242045	0.205368	0.284413	0.257619
75%	1.0	0.666667	270.000000	0.161680	3456.750000	0.264925	0.235294	0.312756	0.306183
max	1.0	0.866667	13207.000000	0.317142	166137.000000	0.368421	0.446602	0.575949	0.509804

Statistiques pour le Cluster 1 : Villes moyennes à grandes

]:

	cluster	ent_0_sal	nb_ent	personne_seule	population_total	pop_enfants_0_20	pop_adultes_20_40	pop_adultes_40_60	pop_adultes_60_80plus
count	1289.0	1289.000000	1289.000000	1289.000000	1289.000000	1289.000000	1289.000000	1289.000000	1289.000000
mean	0.0	0.758511	54.007758	0.091073	1036.586501	0.287119	0.222058	0.292944	0.197879
std	0.0	0.091763	67.945996	0.029551	1214.164423	0.034097	0.049974	0.054684	0.050189
min	0.0	0.500000	1.000000	0.000000	20.000000	0.173913	0.000000	0.000000	0.000000
25%	0.0	0.696970	17.000000	0.071429	393.000000	0.264901	0.191781	0.264957	0.164537
50%	0.0	0.750000	34.000000	0.086572	733.000000	0.285057	0.219498	0.293785	0.197183
75%	0.0	0.809524	64.000000	0.105672	1252.000000	0.306122	0.250000	0.326693	0.228058
max	0.0	1.000000	750.000000	0.333333	18335.000000	0.452174	0.500000	0.571429	0.454545

Statistiques pour le Cluster 2 : Petites Villes

	cluster	ent_0_sal	nb_ent	personne_seule	population_total	pop_enfants_0_20	pop_adultes_20_40	pop_adultes_40_60	pop_adultes_60_80plus
count	416.0	416.000000	416.000000	416.000000	416.000000	416.000000	416.000000	416.000000	416.000000
mean	2.0	0.836374	27.766827	0.188609	368.307692	0.146859	0.142638	0.286399	0.424104
std	0.0	0.141536	134.221256	0.077849	1541.051211	0.069653	0.081069	0.103026	0.126805
min	2.0	0.250000	1.000000	0.000000	7.000000	0.000000	0.000000	0.000000	0.000000
25%	2.0	0.733333	5.000000	0.144740	81.750000	0.111343	0.095925	0.225368	0.346411
50%	2.0	0.833333	9.000000	0.176471	140.000000	0.154898	0.141315	0.284161	0.419321
75%	2.0	1.000000	17.000000	0.215062	269.250000	0.188522	0.180419	0.336314	0.485758
max	2.0	1.000000	2413.000000	1.000000	24585.000000	0.523810	0.666667	1.000000	1.000000

On remarquera qu'il s'agit du segment de villes **avec près de 20% de la population vivant seule (soit presque 5 points de plus que pour les autres cluster**s) , et plus de **42% de la population âgée de plus de 60** ans **(entre 15 et 20 points de plus que pour les autres clusters)**. On peut imaginer des villes éloignées des grandes agglomérations, avec une part importante de la population de plus de 60 ans, vivant seule. En extrapolant, une population qui est restée implantée dans ces petites villes.

Statistiques pour le Cluster 3 : Villes moyennes à petites

:

	cluster	ent_0_sal	nb_ent	personne_seule	population_total	pop_enfants_0_20	pop_adultes_20_40	pop_adultes_40_60	pop_adultes_60_80plus
count	1140.0	1140.000000	1140.000000	1140.000000	1140.000000	1140.000000	1140.000000	1140.000000	1140.000000
mean	3.0	0.854373	30.386842	0.125545	540.071053	0.232750	0.187092	0.297696	0.282462
std	0.0	0.094503	60.948750	0.040659	702.984735	0.044186	0.060979	0.073021	0.072807
min	3.0	0.650000	1.000000	0.000000	20.000000	0.000000	0.000000	0.000000	0.000000
25%	3.0	0.777778	8.000000	0.097976	171.000000	0.207547	0.150789	0.258065	0.233031
50%	3.0	0.844300	16.000000	0.120011	321.000000	0.236050	0.185240	0.297733	0.278648
75%	3.0	0.923077	31.000000	0.147368	620.500000	0.260519	0.221102	0.336668	0.326338
max	3.0	1.000000	1235.000000	0.400000	6578.000000	0.400000	0.500000	0.600000	0.636364

On visualise bien qu'autre de chaque grande ville, et de façon encore plus marquée autour de Lyon et sur l'axe Lyon Grenoble, on a une « première » couronne de grandes villes, où sont concentrées et la population et un nombre important d'entreprises, avec une répartition de la population par âge assez équilibrée.

Puis une couronne de villes Moyennes / Grandes,

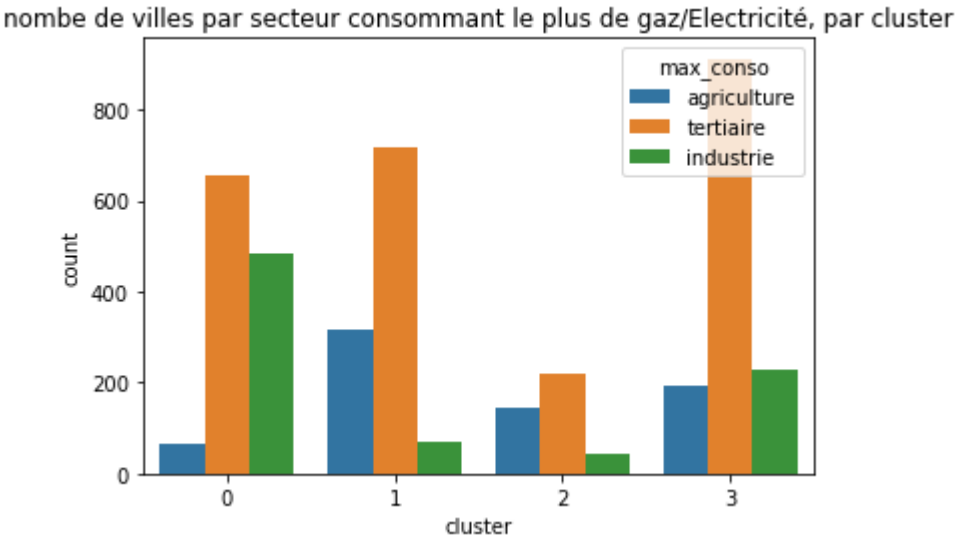
Une couronne de villes Moyennes /Petites,
La part des personnes vivant seules dans ces 3 profils de communes est comprises entre 9% et 13%, en moyenne.

Et enfin en toute périphérie, on trouve des « Petites Villes » avec une population moins importante, et surtout, en moyenne 18% de la population vivant seule, et 42% de la population âgée de plus de 60 ans.

Enfin, en reprenant la notion de « conso_max » définie dans les premiers chapitres (secteur d'activité de la commune ayant la consommation de gaz/électricité la plus élevée), on note que :

- Les villes dont l'activité tertiaire est prédominante sont majoritaires pour tous les Clusters
- Elles sont encore plus dominantes dans le cluster 3 (villes moyennes à grandes)

- Pour le Cluster 2, celui des petites villes avec près de 40% de la population de plus de 60 ans et 20 % de la population vivant seule, on note que 50 % des villes sont celles où l'activité prédominante (conso gaz / électricité) est agricole (environ 200 sur 400 communes), alors qu'elles ne représentent que 10 à 25% des communes sur les autres clusters.



5.1.6 Conclusion

On visualise grâce à la « clusterisation » des communes situées dans un périmètre de 150 km autour de Lyon, une segmentation des communes en fonction de leur éloignement d'une grande ville.

Lyon est très nettement un pôle avec 4 différentes « strates » de territoires.
A proximité, dans la toute proche couronne, des Grandes villes, fortement peuplées, avec des consommations de gaz et d'électricité prépondérantes pour les secteurs Tertiaires et Industriels.
Dans la couronne la plus éloignée, on trouve des petites villes, voire des villages, avec une population bien moins nombreuse, composée pour 40% de personnes de 60 ans et plus, avec une activité Agricole prépondérante.

Ces communes dites rurales sont situées à la jonction des grandes périphéries entre deux grandes agglomérations.

On note le même phénomène autour de Clermont-Ferrand, ville bien moins importante que Lyon, mais capitale pour l'Auvergne.

5.2 Calcul des Clusters sur les communes situées dans un périmètre de 140 km autour de Lyon – avec la variable Salaire

5.2.1 Pré-traitement

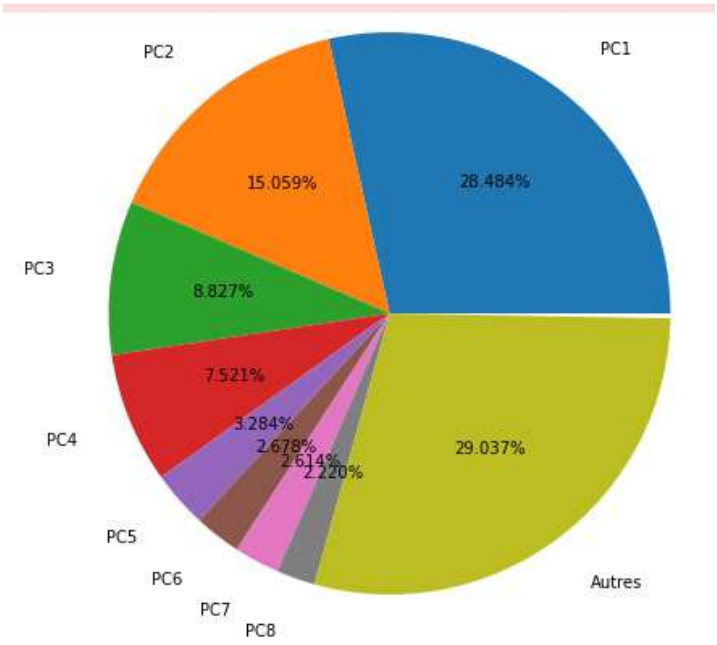
Nous appliquons sur le jeu de données master_salary la même démarche appliquée dans le chapitre précèdent et les mêmes pré-traitements

Le jeu de données df_lyon_num issue des retraitements contient 660 communes à 150 km de lyon et 65 variables numériques.

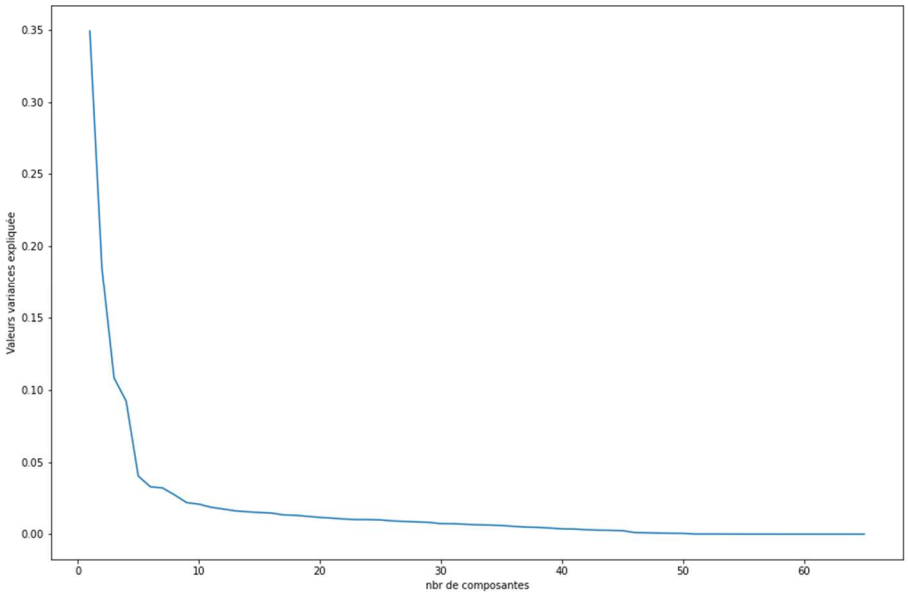
5.2.2 Mise en œuvre de l'Analyse des Composantes Principales (PCA)

⇨ Variance expliquée et nombre de composantes principales

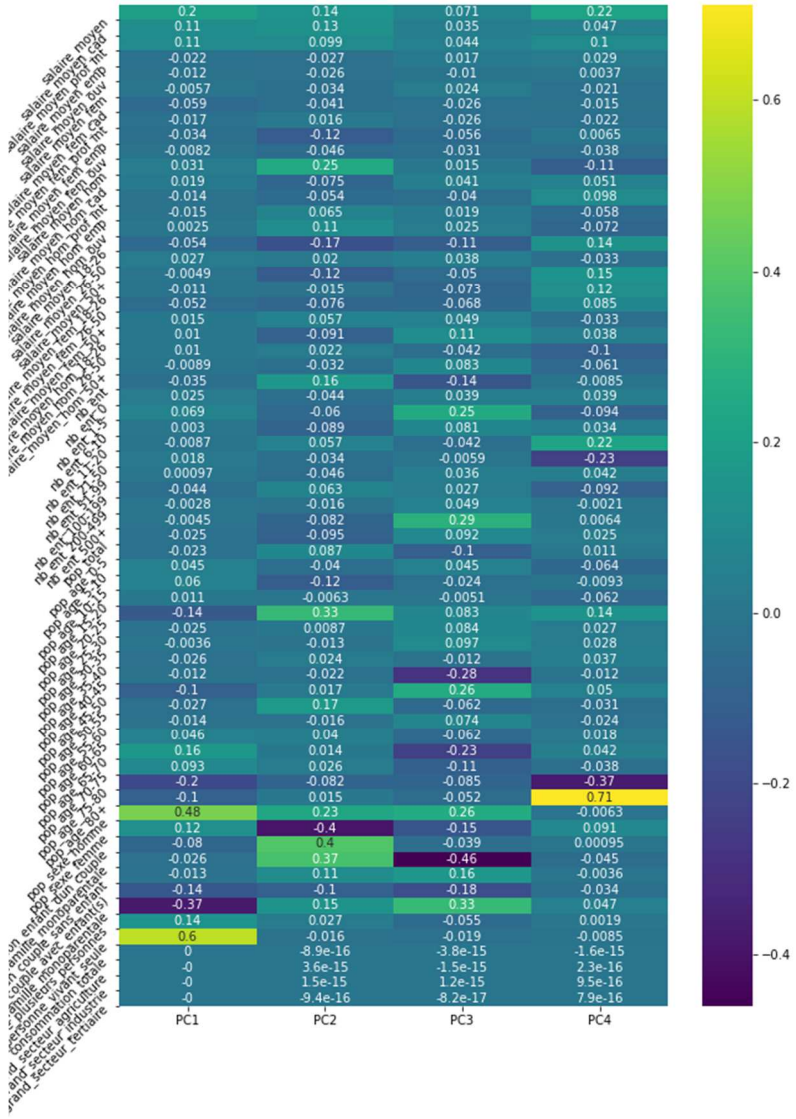
Le camembert suivant montre que ce n'est qu' à partir de la 4ème composante principale que la variance expliquée atteint 59% et il faut aller jusqu'à 8 composantes principales pour atteindre 70% de la variance expliquée.



La courbe suivante montre aussi que le 1er coude important est constaté au niveau de la 4ème composante puis à partir de la 8ème composante, le rythme de l'augmentation devient très faible et, donc, les composantes suivantes ne rapporte que très peu d'explication.


Valeurs variances expliquée
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
10
20
30
40
50
60
nbr de composantes

⇨ Corrélation des 4 premières composantes principales avec les 65 variables



pop_mode_cohabitation_personne_vivant_seule est la variable plus corrélées avec la 1er composante principales puis au 2ème niveau on distingue les variables pop-sexe_homme et pop_mode_cohabitation_adulte_dune_famille_monoparentale (corrélation négative)

Les variables pop_sexe_femme et pop_mode_cohabitation_enfant_dun_couple sont les plus corrélées avec la 2ème composante

La pop_mode_cohabitation_enfant_dune_famille_monoparentale est la plus corrélées avec la 3ème composante.

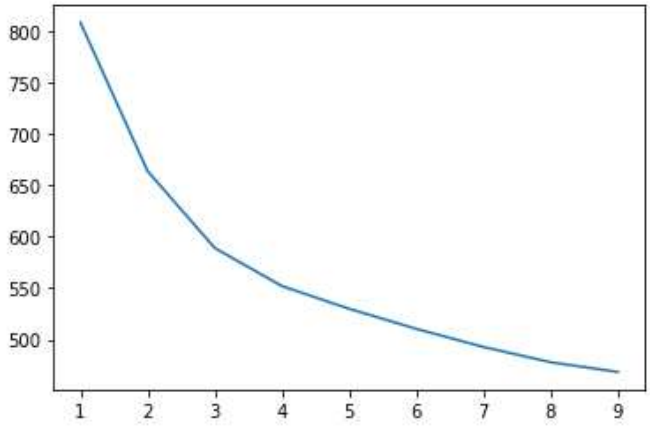
Ensuite on distingue une corrélation significativement forte de la variable pop_age_80+ avec la 4ème composante.

5.2.3 Mise en œuvre du Modèle de Clustering KMeans

⇨ Création du jeu de données avec les composantes principales calculées précédemment

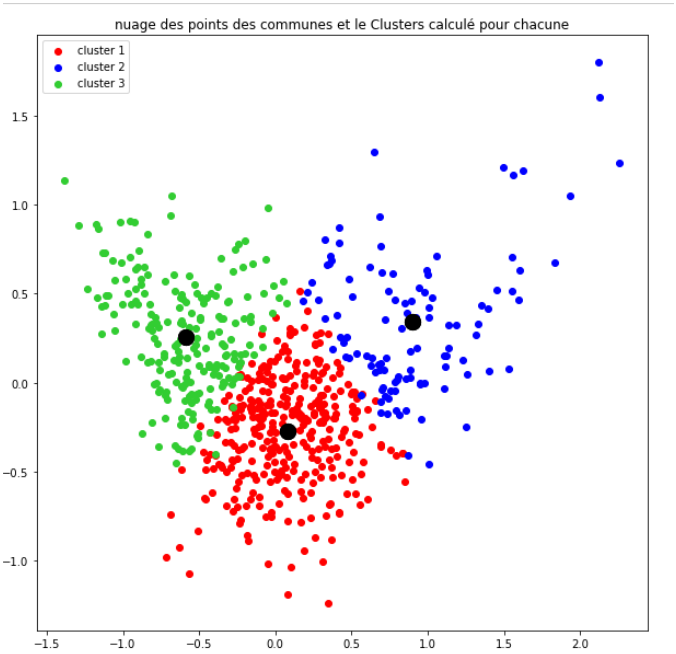
```
n_variable = 65
pca_lyon_kmeans = pd.DataFrame({"Axe " + str(i + 1): coord[:,i] for i in range(n_variable)})
```

⇨ Inertia et méthode du coude pour déterminer le nombre idéal des clusters (classes)

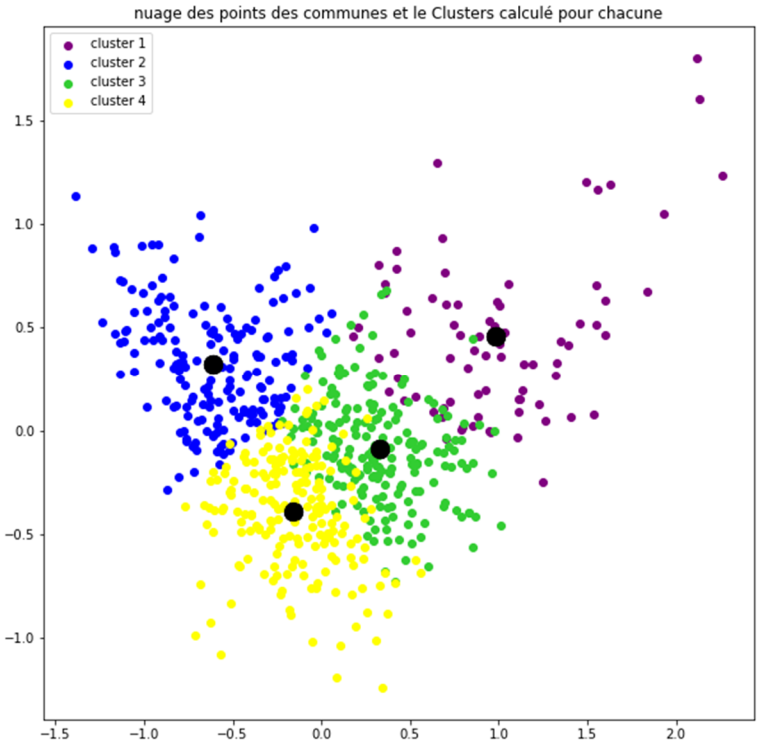


Le nombre de clusters à considérer semble compris entre 3 et 4.

⇨ Application de KMeans sur le dataframe pca_lyon_kmeans
Présentation graphique ave 3 clusters

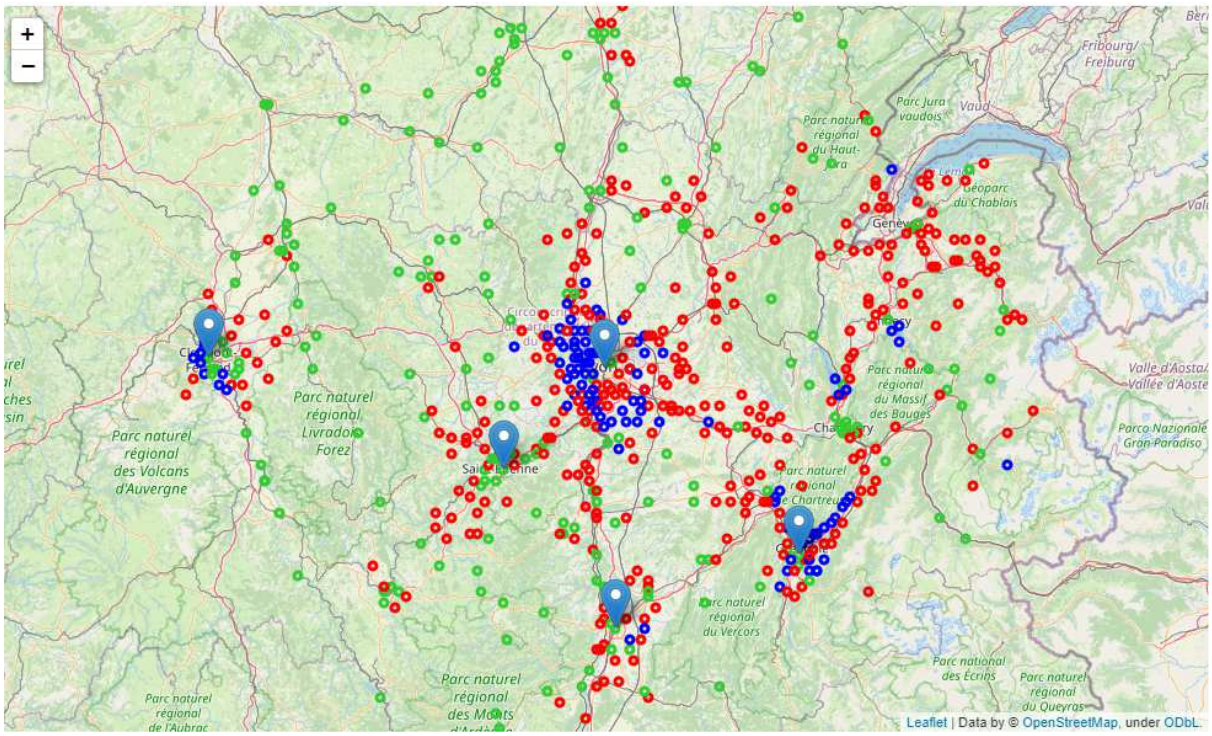


Présentation graphique ave 4 clusters



Le modèle avec 3 clusters affiche une meilleure séparation entre les différents clusters. Les centroïdes semblent mieux positionnés quand on considère 3 clusters.

5.2.4 Visualisation des communes classifiées en 3 clusters, dans un périmètre de 140 km autour de Lyon



5.2.5 Analyse des clusters

Les moyennes conditionnelles

	Axe 1	Axe 2	Axe 3	Axe 4	Axe 5	Axe 6	Axe 7	Axe 8	Axe 9	Axe 10	Axe 11	Axe 12	Axe 13	Axe 14	Axe 15	Axe 16	Axe 17	Axe 18	Axe 19	Axe 20	Axe 21	Axe 22	Axe 23	Axe 24	Axe 25	Ax
0	0.08	-0.27	-0.06	0.02	0.01	-0.01	-0.00	0.0	-0.00	-0.00	-0.00	0.00	0.00	0.01	0.01	-0.00	0.00	-0.00	-0.01	0.0	-0.0	0.0	-0.0	0.0	-0.00	-0
1	0.90	0.34	0.10	-0.03	-0.01	0.01	-0.01	-0.0	0.01	0.01	-0.01	-0.01	-0.01	-0.02	-0.01	-0.01	-0.01	0.00	0.01	0.0	0.0	0.0	0.0	-0.0	0.01	0
2	-0.59	0.26	0.04	-0.01	-0.00	0.01	0.01	-0.0	0.00	0.00	0.00	-0.00	0.00	-0.01	-0.01	0.01	0.00	0.01	0.01	-0.0	0.0	-0.0	0.0	0.0	-0.00	0

Cluster 1

L'axe 2 semble celui qui détermine plus le cluster 1 mais il est difficile de déduire quelles sont les variables influentes.
Sur la carte (les points rouges), on distingue bien ces communes qui se trouvent principalement dans les grandes agglomérations et à travers les axes les importants entre les grandes villes.

Cluster 2
L'axe 1 détermine largement le cluster 2
La variable "les personnes vivants seules" est la plus corrélée avec l'axe 1,
On peut déduire que le cluster 1 est déterminé en majorité par les personnes vivant seules.
On trouve ce mode de vie plus dans les grandes villes et les grandes agglomérations ((les points bleu sur la carte).

Cluster 3
le cluster 3 est inversement influencé par l'axe 1 qui est aussi inversement influencé par la variable pop_mode_cohabitation_adulte_dune_famille_monoparentale.
On peut déduire que le mode de famille monoparentale est probablement plus présent dans ce cluster.
d'ailleurs l'axe 2 influence aussi ce cluster à travers l'autre mode de famille mono parentale (la variable pop_mode_cohabitation_enfant_dune_famille_monoparentale)
cet analyse est probablement confirmé aussi par les points verts sur la carte qui montrent des communes se situant en général loin des grandes villes et agglomérations.

5.2.6 Conclusion

L’impact du salaire moyen est quasiment nul sur l’analyse des clusters et en plus ce jeu de données a neutralisé l’impact de la consommation, probablement, à cause du nombre très réduit des communes.

6 Bilan

Dans le cadre de ce projet, les jeux de données et le nombre de variables disponibles pour les communes en France sont très nombreux.
L'objectif initial décrit dans la fiche projet est assez large.
Aussi, l'une des principales difficultés était de définir les axes d'analyse et de bien les garder en tête tout au long de l'exploration des données. Et ainsi de sélectionner les variables pertinentes pour notre analyse.
La seconde difficulté était relative au contenu des jeux de données. Si les données fournies par l'Insee sont très détaillées, des informations restent manquantes : absence de données de géolocalisation pour les DOM-TOM, données sur le salaire disponibles sur un nombre restreint de communes.... Aussi, le "scope" choisi a dû être ajusté en fonction de l'analyse à produire.

La France est composée de territoires très variés, à multiples visages, soit en fonction de la proximité avec une métropole, soit en fonction de l'emplacement géographique du territoire, au niveau Macro.
Aussi, il était intéressant de faire un focus sur une grande ville et de tenter la mise en œuvre d'un modèle de Machine Learning :
Un modèle de Clustering semblant le plus indiqué pour faire éventuellement ressortir différents profils de villes et les critères de segmentation retenus. Notre hypothèse de travail a pu être vérifiée. Au fur et à mesure que l'on s'éloigne d'une grande ville, on trouve des communes avec des profils différents, jusqu'à se rapprocher à nouveau d'une autre grande ville.