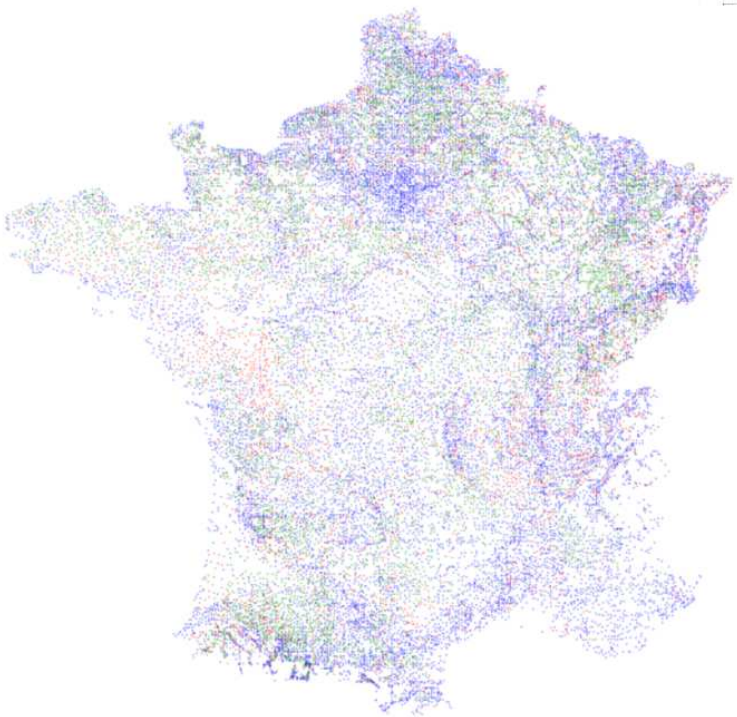


Projet French Industry

Promotion DA Continue – Juin 2021

Anne Schneider
Omar El Ghazi

DataScientest

Objectif : Comparer les inégalités en France

A partir de bases **de données de l'INSEE,** indiquant pour chaque ville française

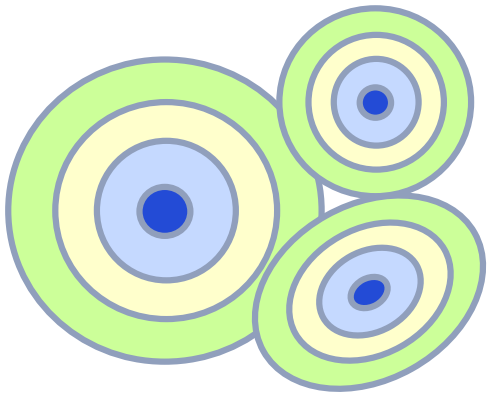
- Données Géographiques (y compris longitude, latitude)
- Le **nombre d'entreprises**, classées par taille
- Les **informations démographiques** (population, âge, genre, mode de vie)
- **Salaires net moyen** par catégorie d'emploi, par genre*

Créateur : Tomas Griger
Droits d'auteur : Tomas Griger

36 840 villes françaises considérées pour ces jeux de données INSEE

**Les salaires nets moyens sont disponibles pour les zones de plus de 2000 habitants (couvre 5 136 communes)*

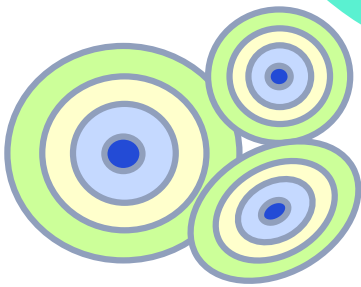
Problématique



- Une aire urbaine est décrite comme composée de 3 espaces différents
  - La ville-centre
  - **La banlieue**
  - La zone périurbaine
- Entre 2 aires urbaines, qu'en est-il des territoires en grande périphérie ?
- Quel est le rapport entre le tissu économique et la répartition de la population, sur les territoires compris entre les grandes agglomérations ?

Hypothèses

- **Dans les grandes villes et les grandes agglomérations**, et à **leur proche périphérie**, on trouve principalement **des entreprises du secteur tertiaire, une concentration d'entreprises de grande taille,** une **population importante** 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 population moindre et **des revenus moins élevés**

- **J**usqu'à se **rapprocher d'une autre grande agglomération.**

Un nouveau jeu de données a été ajouté :
Consommation annuelle d'électricité et de gaz par commune et code NAF
Il va permettre de :

- *Introduire les **secteurs d'activité** dans le projet, grâce aux consommations Gaz/Electricité relevé par commune,*
- *regroupés pour les **Grands Secteurs Tertiaires, Industriels, Agricoles**, sur la base du code NAF.*

Définition :
Le **secteur d'activités ayant la plus forte consommation Gaz/Electricité** sur la commune est considéré comme le **secteur d'activités prépondérant de la commune**

Jeux de données – Traitement et Fusion

- **Un jeu de données complémentaire** :

Consommation annuelle d'électricité et gaz par commune, code NAF, et secteur d'activité

- Consommation des Secteurs Tertiaires, Industriels, Agricoles sur la commune

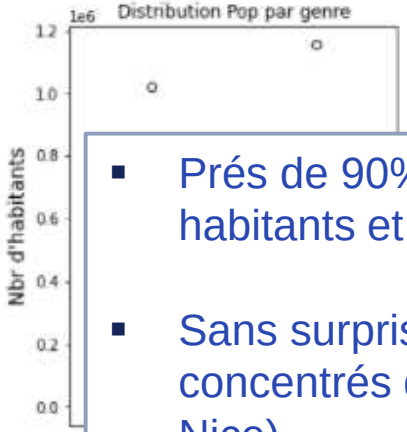
- Variable max-conso : Le secteur d'activités avec la plus forte consommation sera considéré comme le secteur d'activité prépondérant de la commune

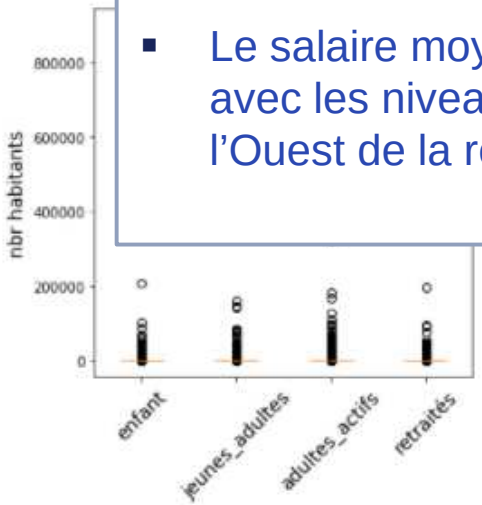
- **Fusion des Jeux de données** : le code commune (code insee) est la clé commune à l'ensemble des jeux, **74 variables résultantes**
- **Le même processus traitement** est appliqué à tous les jeux, moins de 10% des entrées sont supprimées, **31 780 communes restantes**

DataScientest

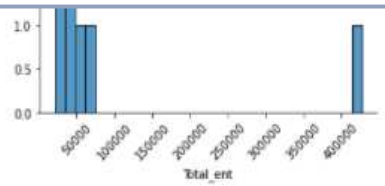
Distributions des variables clés

Population

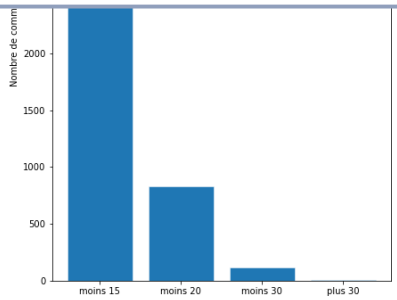




Entreprises



Salaires



- Prés de 90% des communes en France ont une population inférieure à 2500 habitants et comptent en moyenne 30 entreprises sur leur territoire
- Sans surprise, les plus fortes populations et le plus grand nombre d'entreprises sont concentrés dans les grandes agglomérations (Paris, Marseille, Lyon, Toulouse, Nice)
- Le salaire moyen net le plus répandu est de moins de 15 euros / h, les communes avec les niveaux de salaires les plus élevés (plus de 30€/h) sont situées dans l'Ouest de la région parisiennes (78, 92).

Quelques Chiffres clés :

98% des Entreprises
ont entre **0 et 5 salariés**

Les 10 plus grandes agglomérations
- **10% de la population**
- **20% des entreprises**

Les femmes à poste équivalent et sur un temps plein ont un salaire 17% inférieur à celui des hommes.

80 % des communes en France comptent moins de 2 000 habitants

24% des familles monoparentales sont concentrées dans les grandes agglomérations

L'âge moyen en France est 42 ans

Dans les villes **de moins de 2000 habitants**, les **plus de 60 ans représentent 18% de la population**,

Dans les villes de plus de **100 000 ils représentent seulement 15% de la population**

En France :
- **54%** des communes ont une activité **Tertiaire** prépondérante**,**
- **29%** une activité **Agricole**
- **17%** une activité **Industrielle**

Dans les communes de + 200 000 habitants,
- **90%** des communes ont une activité **Tertiaire** prépondérante**,**
- **10%** une activité **Industrielle**

DataScientest

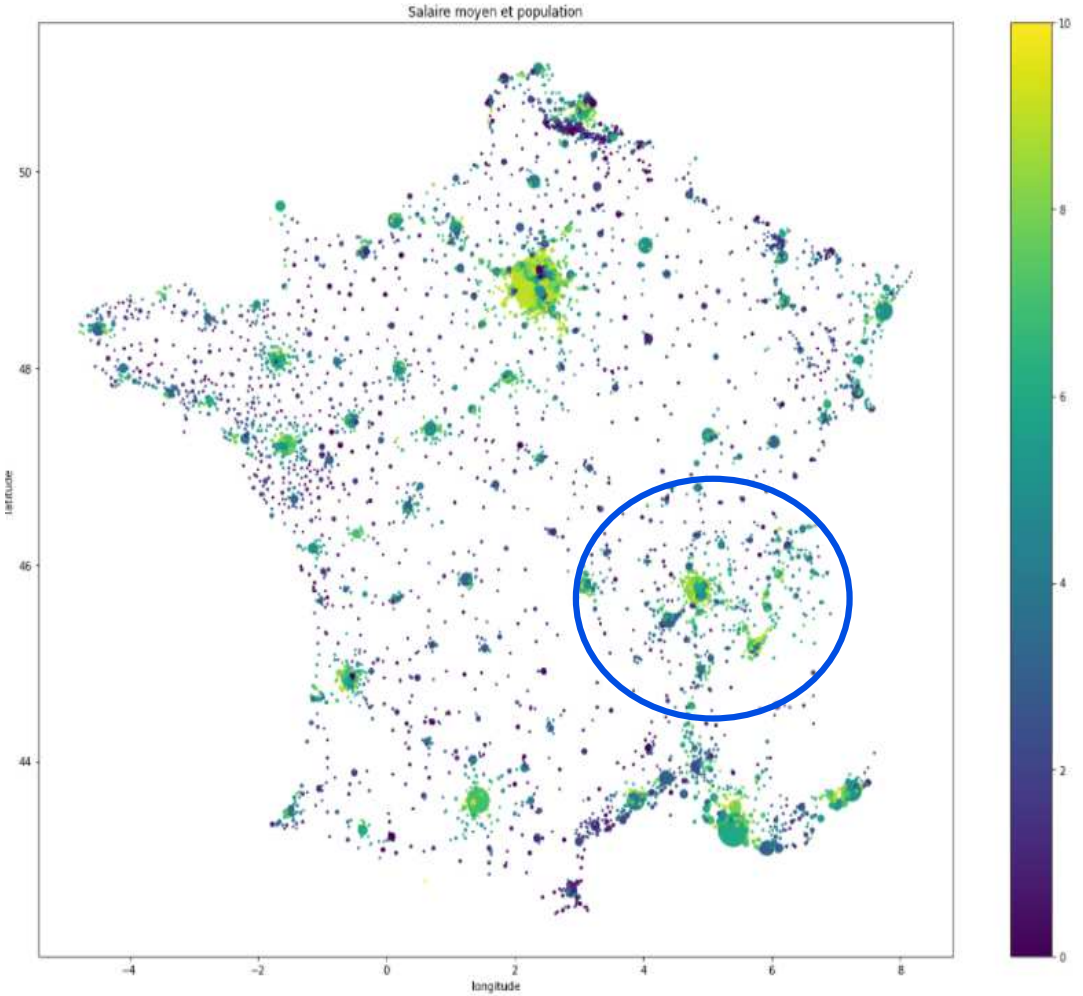
Population et salaires en France

- **Taille des points : en fonction la taille population**, plus la taille est grande plus la population est importante dans la commune.
- **La couleur des points** : en **fonction du salaire** horaire net moyen, plus la couleur est claire plus le salaire moyen est élevé.

Constat : Les salaires nets moyens sont plus élevés dans les grandes agglomérations, qui sont également les plus peuplées.

Largement attribués aux effets « de composition » et « de densité » de l'activité économique, les écarts tiennent aussi à la concentration de personnes très diplômées. (source : strategie.gouv.fr)

Composition : poids relatif de certaines activités économiques et de certaines catégories socio-professionnelles dans une zone d'emploi

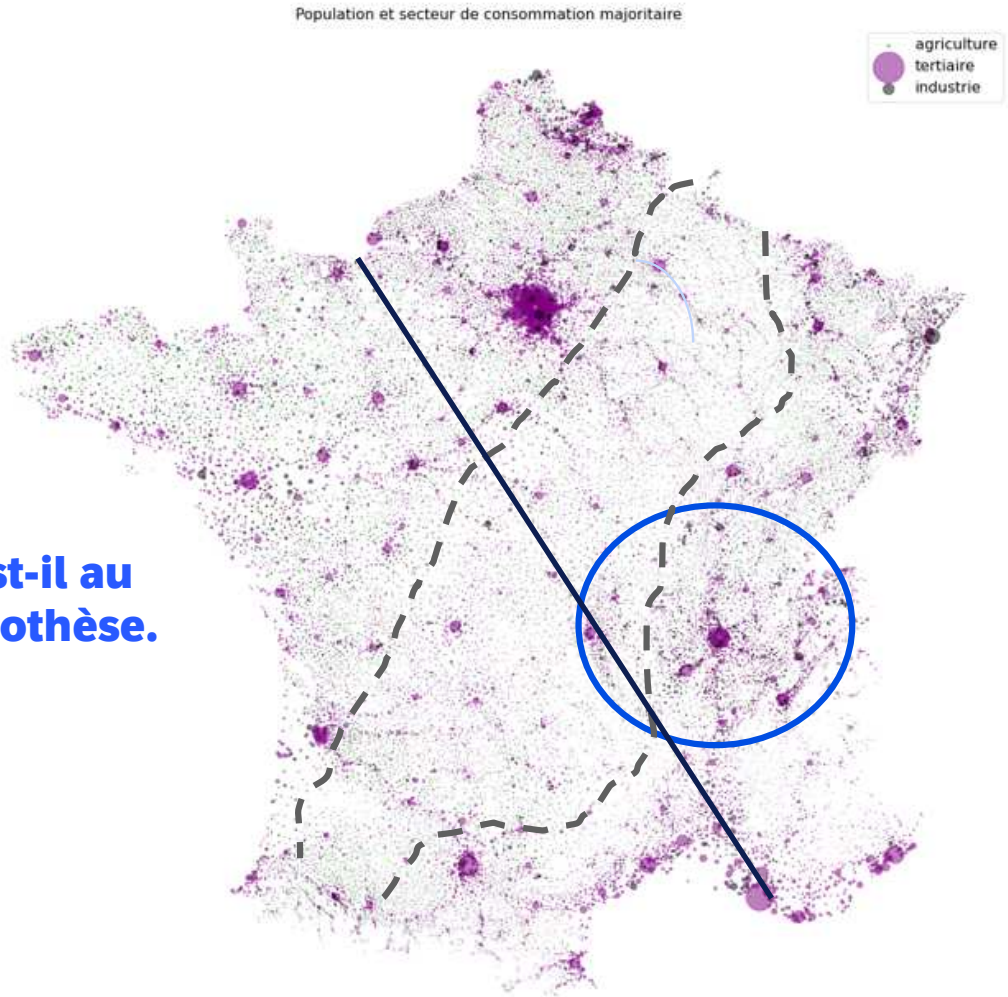


DataScientest

Population et activité en France

- **Taille des points : en fonction la taille population**,
- **La couleur des points** : en fonction du secteur d'activité prépondérant sur la commune

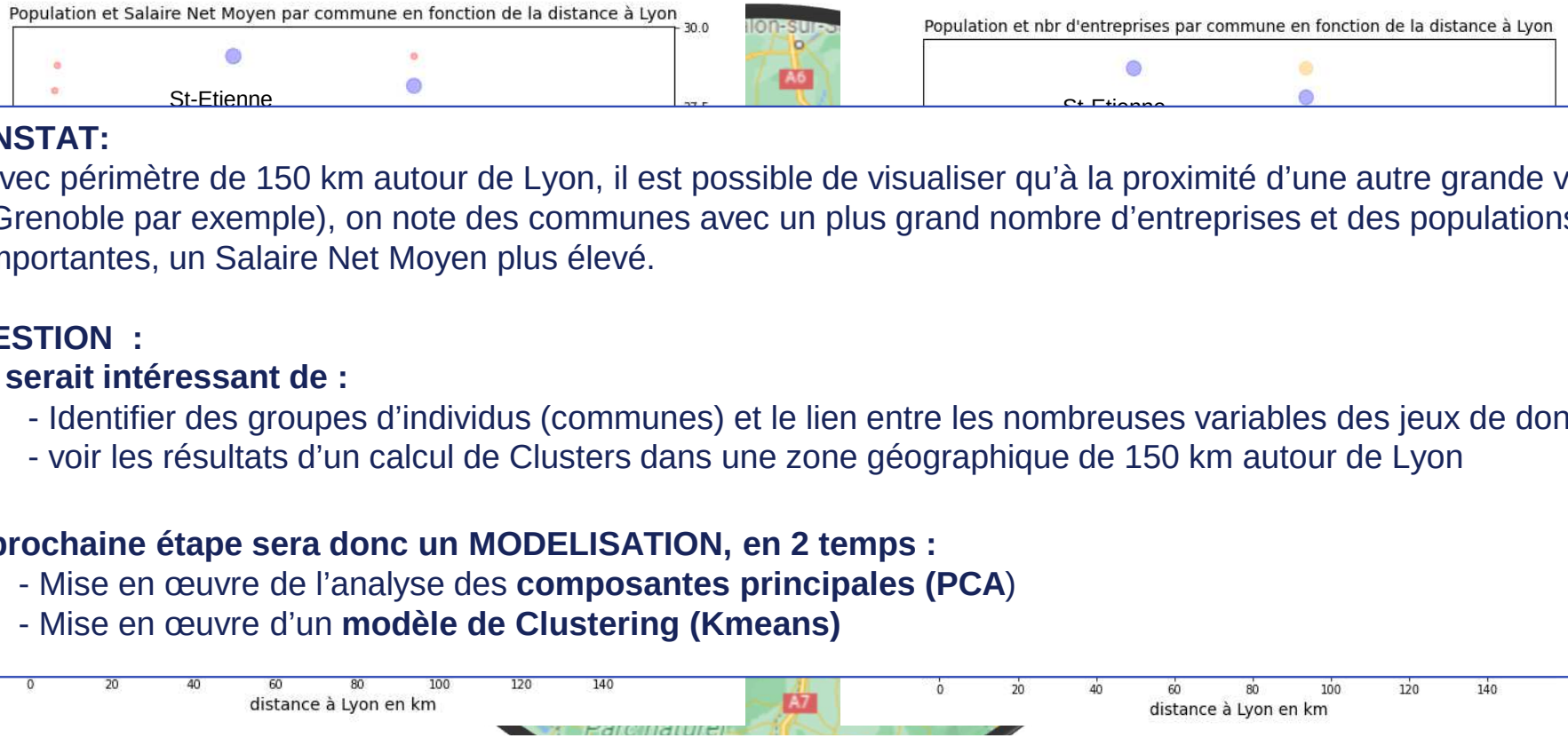
Constat : Les grandes agglomérations et leur proche périphérie sont les zones les plus peuplées, et avec une activité Tertiaire dominantes, et quelques pôles Industriels.
Les territoires éloignés des agglomérations sont moins peuplés. Les activités dominantes sont plutôt agricoles, puis tertiaires et industrielles.



Fort de ces 2 constats au niveau National, qu'en est-il au niveau Régional. Pouvons-nous vérifier notre hypothèse. Région choisie : Lyon et sa très grande périphérie.

DataScientest

Focus sur la Lyon et sa très grande périphérie



CONSTAT:

- Avec périmètre de 150 km autour de Lyon, il est possible de visualiser qu'à la proximité d'une autre grande ville (Grenoble par exemple), on note des communes avec un plus grand nombre d'entreprises et des populations plus importantes, un Salaire Net Moyen plus élevé.

QUESTION :

- **il serait intéressant de :**
  - Identifier des groupes d'individus (communes) et le lien entre les nombreuses variables des jeux de données
  - voir les résultats d'un calcul de Clusters dans une zone géographique de 150 km autour de Lyon

La prochaine étape sera donc un MODELISATION, en 2 temps :

- Mise en œuvre de l'analyse des **composantes principales (PCA**)
- Mise en œuvre d'un **modèle de Clustering (Kmeans)**

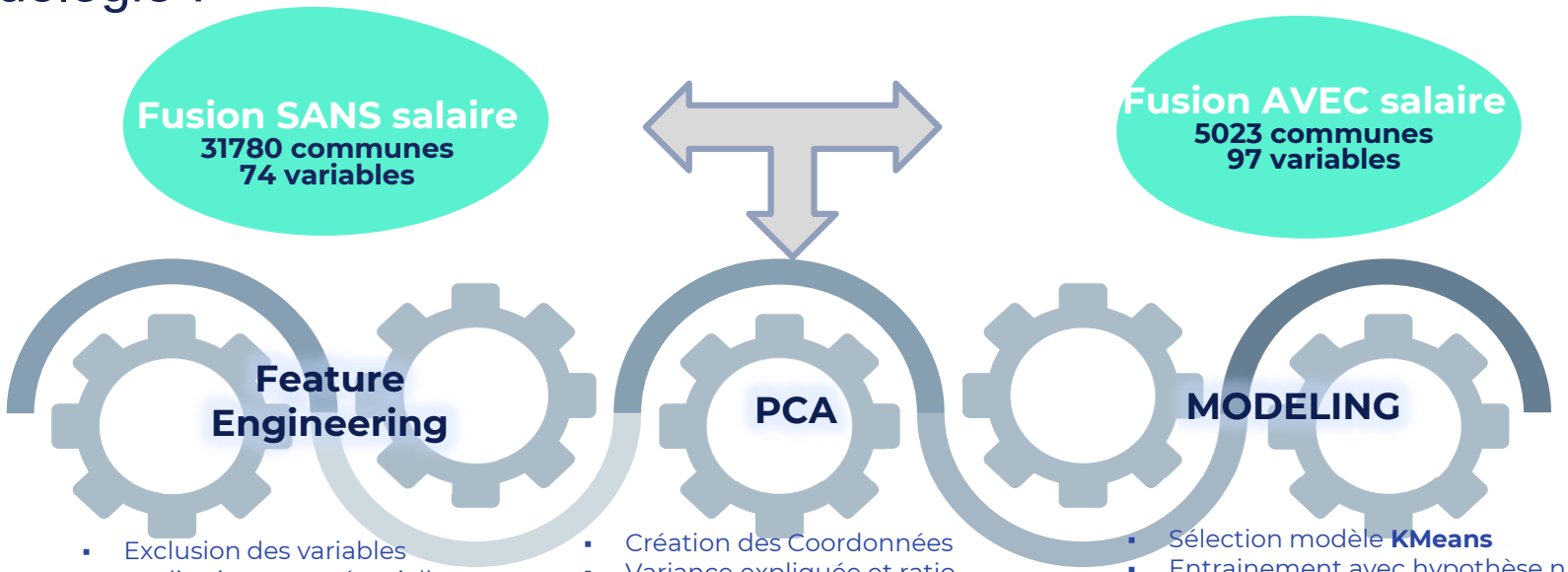
- On note clairement un « **effet Grande Ville** » sur le **Salaire Net Moyen**, avec une effet moindre pour Saint-Etienne.
- A partir d'un rayon situé à 30 km de Lyon, **et tous les 20 km, on trouve des communes dont la population est comprise entre 25 000 et 50 000 habitants**

Modélisation

- **Question** :

Peut-on segmenter les communes dans un périmètre de 150 km autour de Lyon ?

- Méthodologie :



- Exclusion des variables qualitatives et catégorielles
- Pondération des sous catégories
- Exclusion des NaN
- Scaling

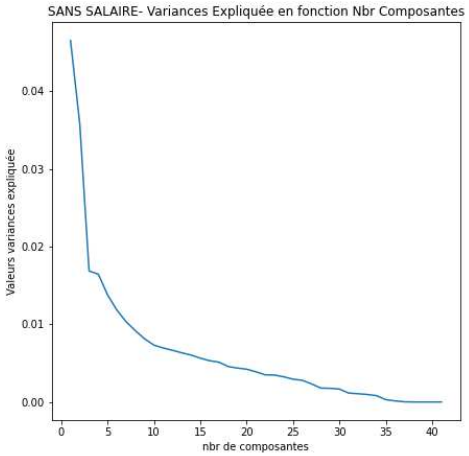
- Création des Coordonnées
- Variance expliquée et ratio nombres de Comp. Principales à considérer
- Corrélation des variables avec les composantes principales

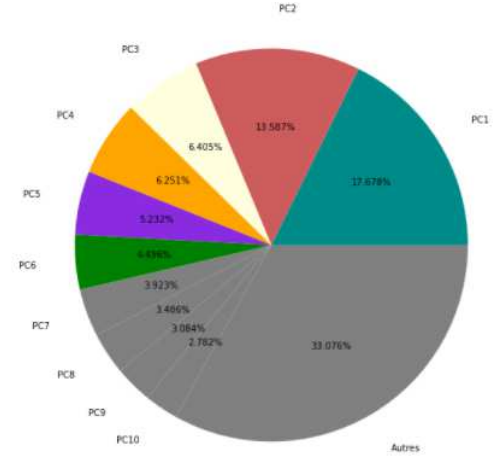
- Sélection modèle **KMeans**
- Entrainement avec hypothèse nbr PC et n_cluster
- Evaluation pour sélection de l'hypothèse optimale (#PC et n_clusters)

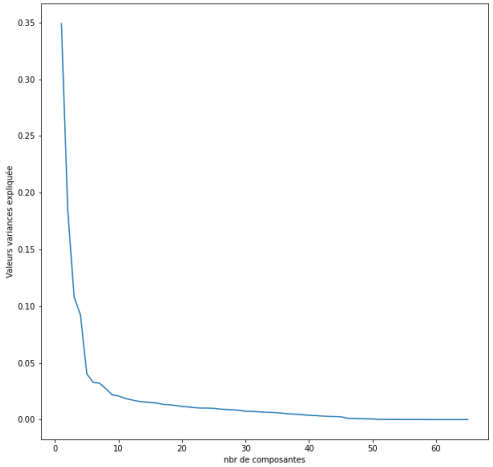
DataScientest

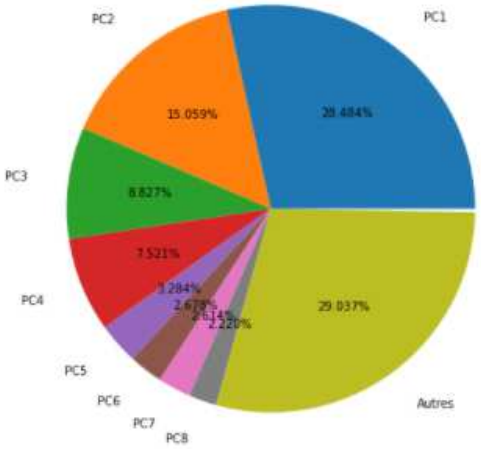
ANALYSE DES COMPOSANTES PRINCIPALES

- Chacun des jeux de données contient respectivement 74 variables (sans salaire) 97 variables (avec salaires) .
- Mise en œuvre de PCA pour obtenir une synthèse de l'information et mieux quantifier les corrélations entre les variables









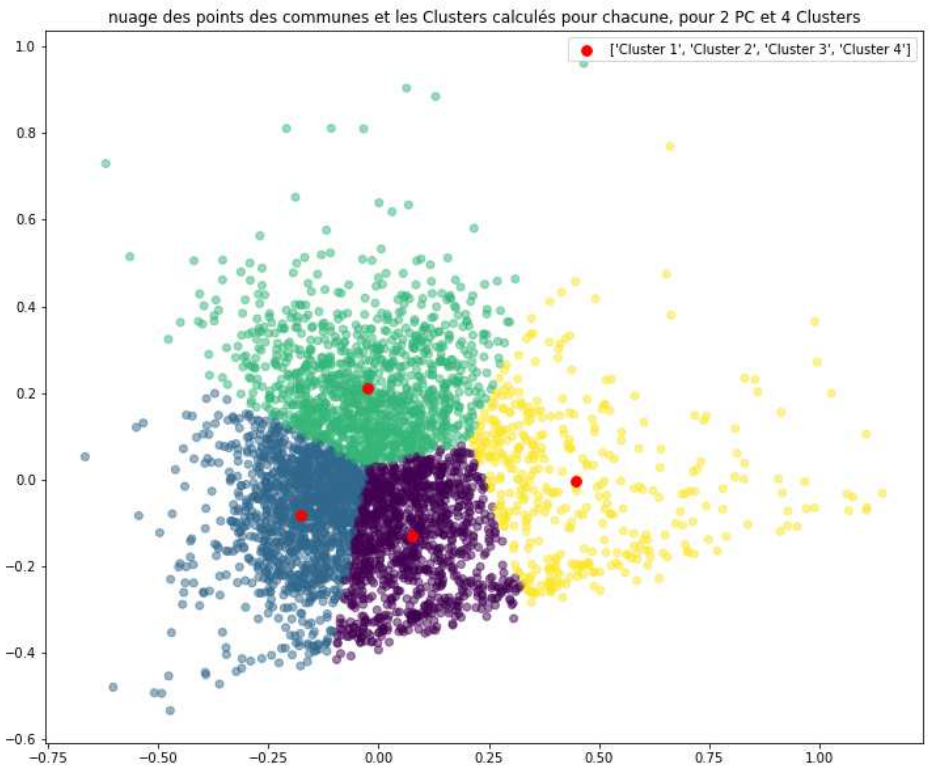
- **Sans Salaire** : On constate un « coude » pour 2 Composantes Principales, PC1 & PC2.
- Elles permettent d'atteindre 30% de variance expliquée. Et sont égales à plus du double du ratio des autres composantes.

- **Avec Salaire :** On constate un coude à 4 Composantes Principales
- Elles permettent d'atteindre plus de 70% de variance expliquée

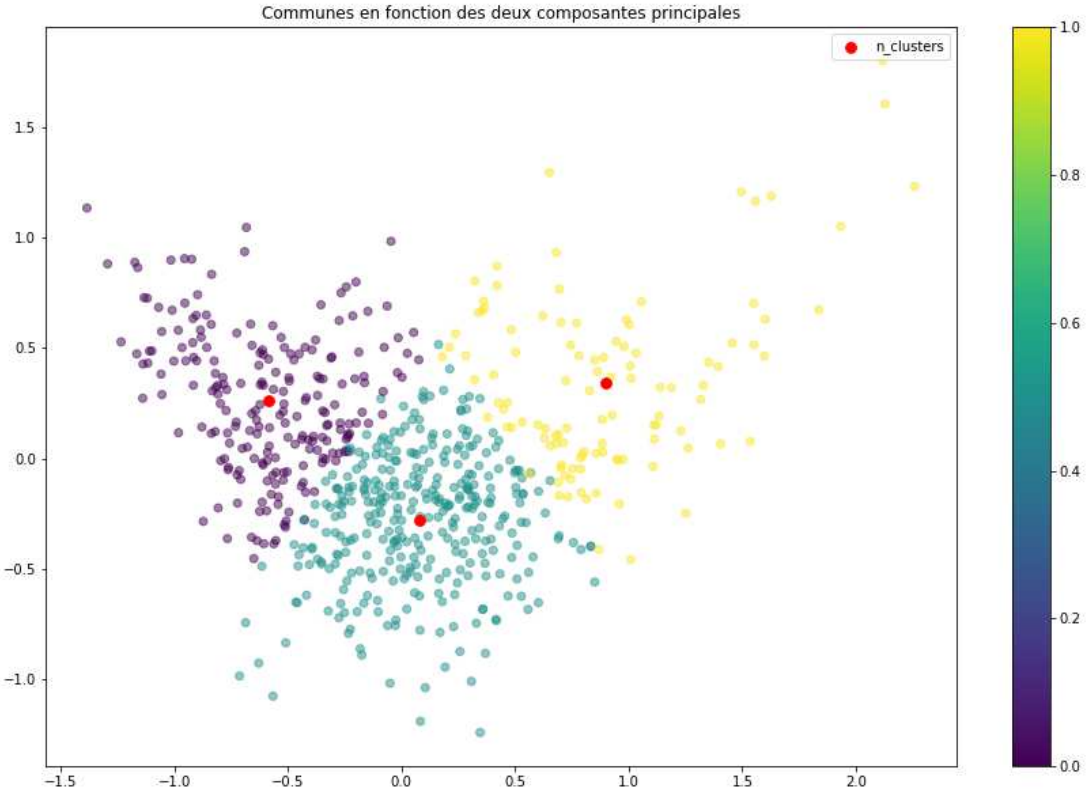
DataScientest

N-clusters Kmeans après évaluation et optimisation

SANS SALAIRE 2 PC, n_clusters = 4



AVEC SALAIRE- visualisation avec 2 PC, n_clusters = 3



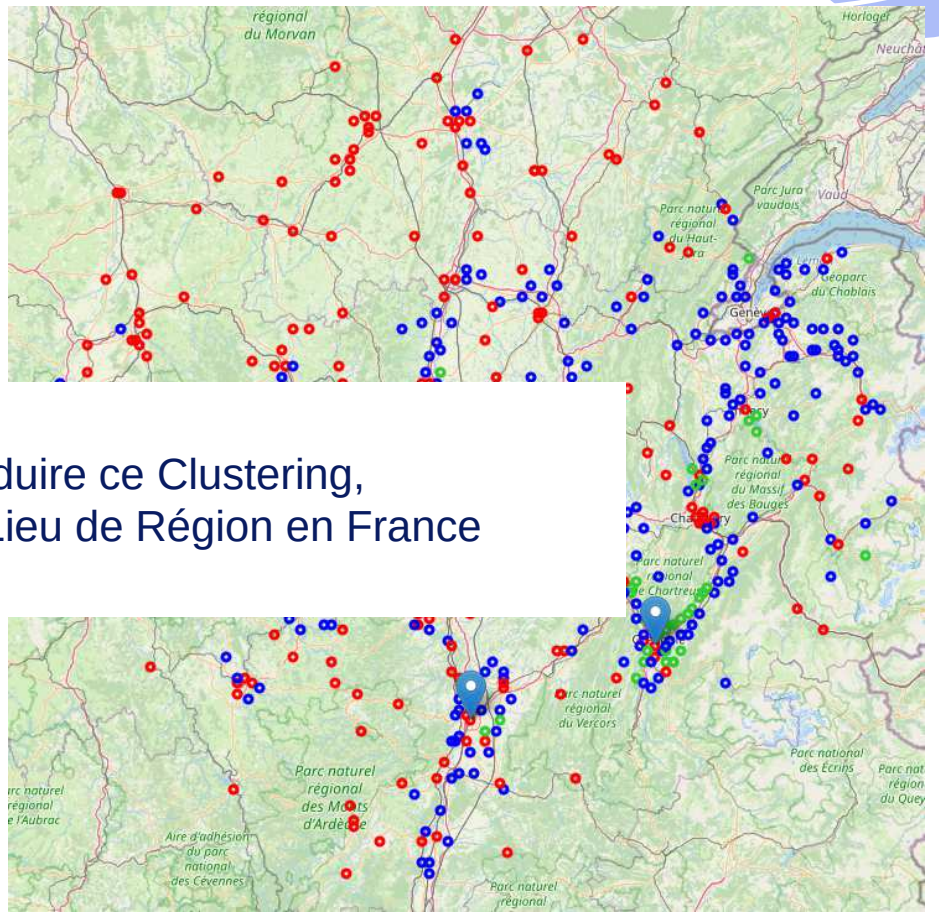
DataScientest

Résultats Modèle Kmeans

Segmentation des villes situées dans un rayon de 150 km autour de Lyon

SANS SALAIRE

AVEC SALAIRE



Next Step :
Une application visant à reproduire ce Clustering,
mais en choisissant un Chief Lieu de Région en France
et la distance au Chef Lieu.

DataScientest

BILAN

- Le projet French Industry comporte plusieurs jeux de données avec un nombre très conséquent de variables,
- Les données concernant les Salaire ne sont pas disponibles pour l'ensemble des communes, contrairement aux autres variables
- Aussi l'enjeu principal a été de déterminer un axe de travail, la problématique à traiter, pour ne pas « se perdre » avec la multitude d'informations disponibles , et être en mesure de toujours se raccrocher à notre objectif
- Un grand MERCI à Paul, qui nous a permis de bien établir notre objectif fil rouge, la problématique à traiter, et appris des bonnes pratiques de bon « Codeur » !

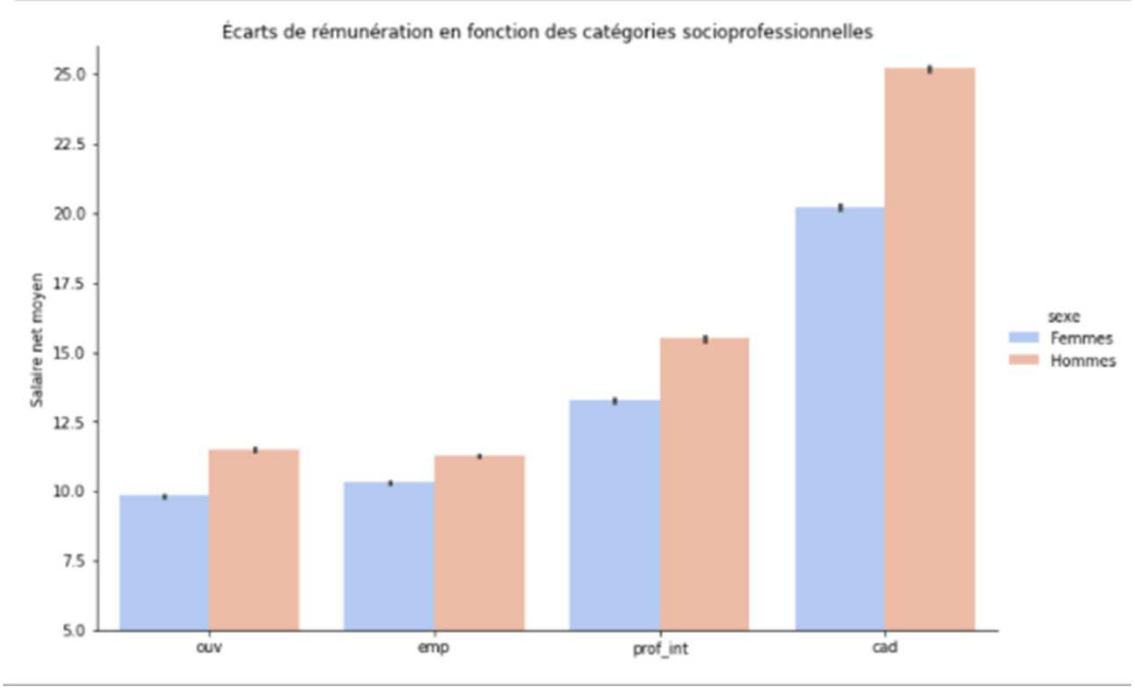
UNE SUPER EQUIPE soutenue, guidée par Paul, notre Super Mentor

DataScientest

Back Up Slides

DataScientest

Focus sur les salaires moyens :



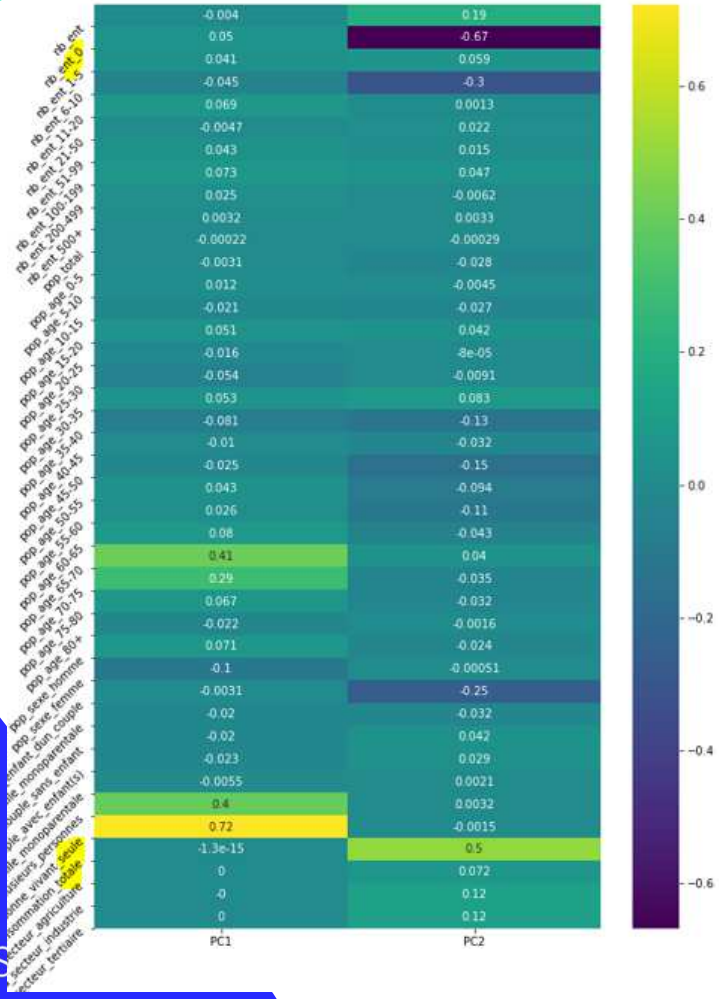
Le **salaire** mensuel net moyen des **femmes en France** est, selon l'INSEE, de **16,8 % inférieur** à celui des **hommes** (pour un temps plein)

- Quelque soit la catégorie d'emploi, les femmes ont un salaire net moyen inférieur à celui hommes,
- L'écart est encore plus marqué pour les Cadres et les dirigeants. Les femmes cadres ont un salaire net moyen 20% inférieur à celui des hommes cadres.

DataScientest

Corrélations des variables aux Composantes principales

SANS SALAIRE 2 PC



AVEC SALAIRE 4 PC