[**DataScientest.com**](http://datascientest.com)

**Projet French Industry**

****

**Par Mathieu LARGAUD et Nicolas DELANGRE**

**Formation Data Analyst**

**Alumni Sep 24 Bootcamp**

**SOMMAIRE**

**01) Préambule**

**02) Rapport d’exploration, de data-visualisation et de pré-processing des données**

1. Rapport d’exploration des jeux de données ……………………………… 4
2. Nettoyage des dataframes et pre-processing ……………………………. 5
3. Data-visualisations et avis métier …..……………………………………… 7
4. Constats sur ces data-visualisations ……………………………………… 14
5. Tests et avis métier …………………………………………………………. 15
6. Constats sur ces tests ……………………………………………………… 16
7. Conclusion sur ce rapport d’exploration ………………………………….. 17

**03) Rapport de modélisation**

1. Première itération de la matrice de corrélation …………………………. 18
2. Seconde itération de la matrice de corrélation …………………………. 20
3. Troisième itération de la matrice de corrélation ………………………… 21
4. Machine Learning et entraînement des modèles ………………………. 22
5. Choix du modèle final ……………………………………………………... 23
6. Analyse des graphiques des erreurs, de répartition, et des résidus …. 23
7. Graphique d’interprétabilité …………………………...………………….. 25
8. Conclusion scientifique et métier ……………………..………………….. 25

**01)** **Préambule**

**Définition de l’INSEE**:

* L’INSEE (Institut National de la Statistique et des Études Économiques) est une direction du ministère de l’Économie et des Finances chargée de la production, de l’analyse et de la publication des données statistiques officielles en France.
* Fondé en 1946, l’INSEE collecte, produit et diffuse des informations sur l’économie et la société françaises.

Nous avons été choisi pour travailler sur le projet **French Industry,** composé de **4 fichiers CSV provenant de l’INSEE** :

* Base des établissements par tranche,
* Base des salaires par catégories sociaux-professionnelles,
* Base de la population, par tranche d’age, de sexe, et de mode de vie,
* Base géographique sur les villes françaises.

L'objectif du projet French Industry est d'analyser et de conclure sur les inégalités en France, de façon factuelle et chiffrée, par le biais de graphiques, de tests, et de prédictions à l’aide de valeurs cibles et de Machine Learning, et enfin de présenter ces analyses et graphiques à l’aide de Streamlit et PowerBI.

Afin de nous familiariser avec ces jeux de données et de nous en faire une première idée, nous allons tout d’abord créer des graphiques simples afin de :

* **Comparer les entreprises en fonction de leur localisation et de leur taille :**
  + Analyser la répartition des entreprises par taille (micros / petites / moyennes / grandes) dans différentes régions et villes françaises.
  + Observer les écarts entre grandes et petites villes ainsi que les différences régionales

* **Analyser la population en fonction du salaire et de la localisation :**
  + Comparer les salaires nets moyens dans les villes en fonction de la catégorie socioprofessionnelle (employés, cadres, cadres sup)
  + Analyser les écarts de salaires entre les hommes et les femmes, en précisant la disparité selon l’âge et la localisation géographique.

* **Faire un focus sur un panel de grandes villes :**
  + Zoom sur un panel de 5 grandes villes pour examiner les dynamiques : entreprise par taille, niveau de salaire, structure démographique ou éducation.
  + Identifier les facteurs spécifiques à ce panel qui pourraient influencer les inégalités observées.

Par la suite, et au fur et à mesure de nos constats, nous créerons de nouveaux graphiques et dirigerons de nouveaux tests afin d’affiner notre analyse, afin de déterminer une ou plusieurs valeurs cibles.

**NB** : Afin de mener à bien ce projet, nous utilisons la plateforme **Colab** de Google, et ses

notebooks, afin de faciliter le travail en équipe et le partage de notre travail.

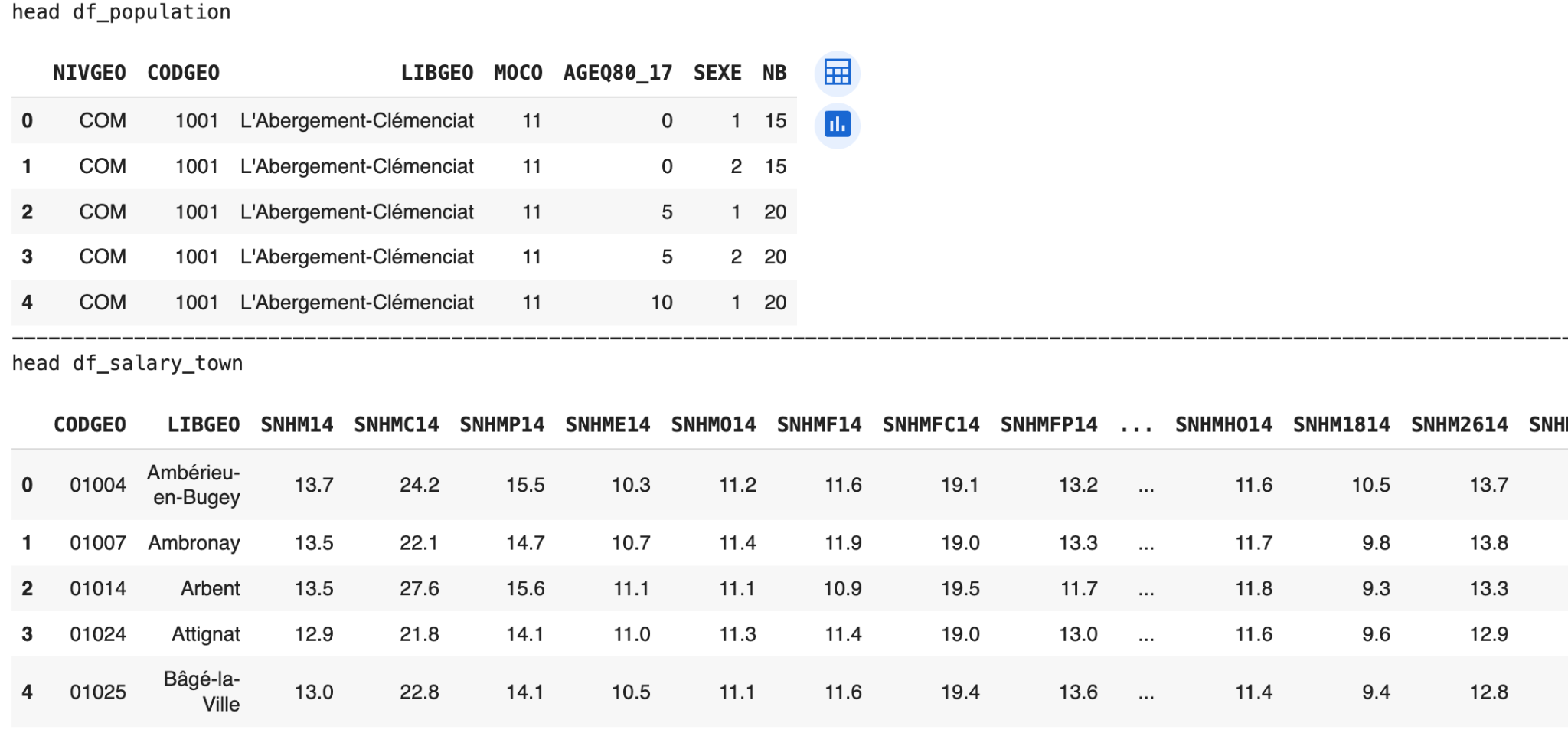
Vous trouverez les liens vers notre dossier Colab et notre Drive Google ci après:

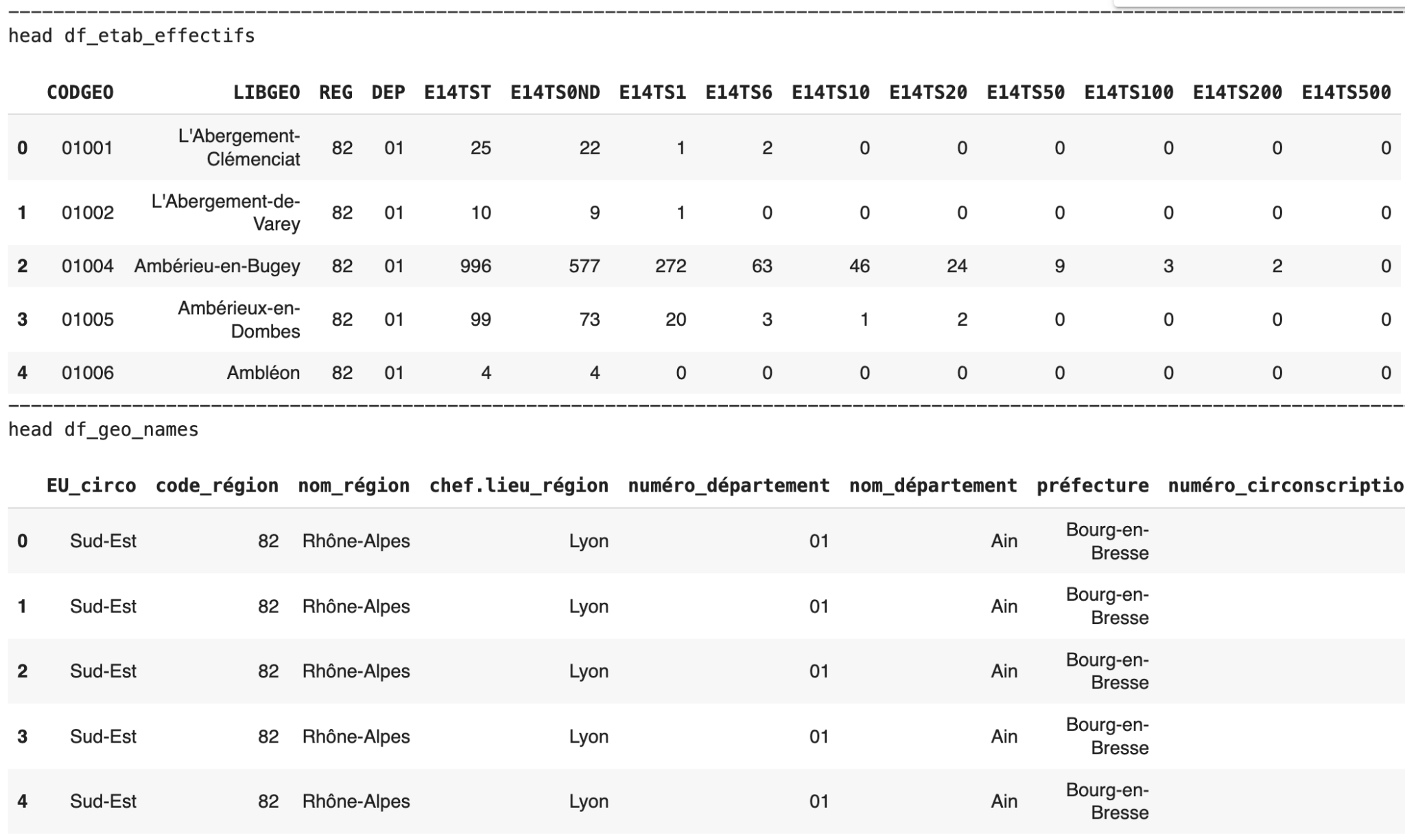
* [Colab Projet\_Industry\_DST Mathieu & Nicolas V2.ipynb](https://colab.research.google.com/drive/1JralA_bx6mv1htDJHvJ4zQKrUjHmZw1K?usp=sharing)
* [Jeux de données](https://drive.google.com/drive/folders/1BlKMS_Gm4EDSEiJBmXIhMsTEmJaVaQaT?usp=drive_link)

**02)** **Rapport d’exploration, de data-visualisation et de pré-processing des données**

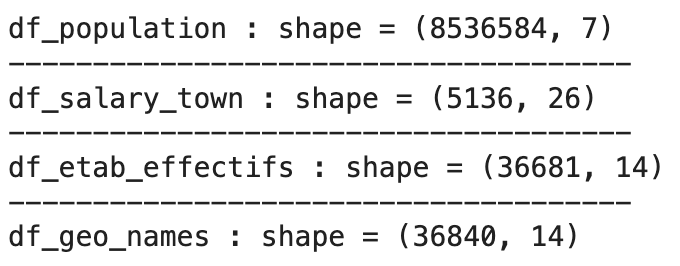
**a)** **Rapport d’exploration des jeux de données**

* Nous avons à notre disposition 4 jeux de données de l’INSEE contenant respectivement des informations relatives à la taille des entreprises, les salaires, la population et enfin des données géographiques (le tout en France):
* Nommage des 4 jeux de données et des df correspondants:
  + base\_etablissement\_par\_tranche\_effectif.csv => **df\_etab\_effectifs**
  + net\_salary\_per\_town\_categories.csv ========> **df\_salary\_town**
  + population.csv =========================> **df\_population**
  + name\_geographic\_information.csv ==========> **df\_geo\_names**
* Nous allons explorer ces dataframes avant d’exposer nos conclusions et d’appliquer nos actions
* Voici un aperçu de chaque DF:

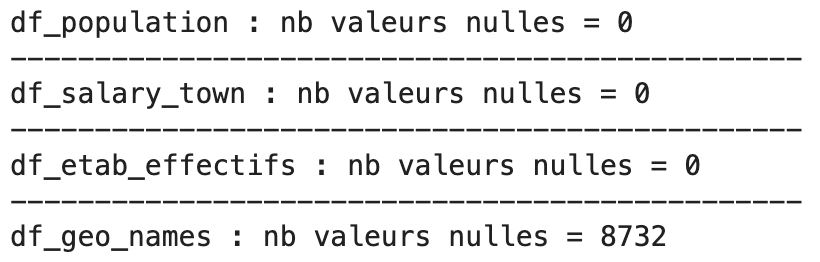




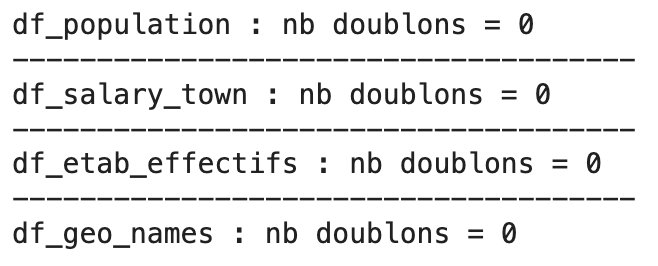
* Voici un aperçu des dimensions de chaque fichier:



* Nous constatons des valeurs manquantes sur le df\_geo\_names:



* Nous ne constatons pas de doublons:



**b)** **Nettoyage des dataframes et pre-processing**

* Sur le df\_etab\_effectifs, nous décidons d’agréger les données tailles pour créer 4 catégories d’entreprises (selon l’exemple de l’INSEE), et de supprimer les 9 autres colonnes relatives à la taille:

# micro entreprise : de 0 à 9 salariés

# petite entreprise : de 10 à 49 salariés

# moyenne entreprise : de 50 à 199 salariés

# grande entreprise : plus de 200 salariés

* Nous constatons que nos 4 dataframes ont une donnée commune avec le CODGEO (qui doit comporter 5 digits):
  + Il faut rajouter un “0” en début de la cellule du CODGEO sur les df\_geo\_names et df\_population
  + Il faut également renommer la colonne “code\_insee” en “CODEGO” sur le df\_geo\_names

Cela va nous permettre d’avoir une homogénéité au niveau du nom de la colonne (CODGEO) et du nombre de digits (5), pour pouvoir merger plus tard nos 4 dataframe en un seul

Nous vérifierons après chaque merge la création d'hypothétiques doublons, et procéderons au drop de colonnes si besoin

* Nous supprimerons sur le df\_geo\_names les colonnes ‘latitude’, ‘longitude’ et ‘éloignement’: Ces colonnes comportent des valeurs manquantes, et sont redondantes avec d’autres données d’autres colonnes tel que le code postal, le département, le nom de la ville …. Cela n’aura aucune influence sur la où les valeurs cibles que nous pourrions choisir plus tard.
* Nous allons également importer un jeux de données de l'INSEE, relatifs aux diplômes, afin de pouvoir affiner nos analyses et établir de nouvelles corrélations

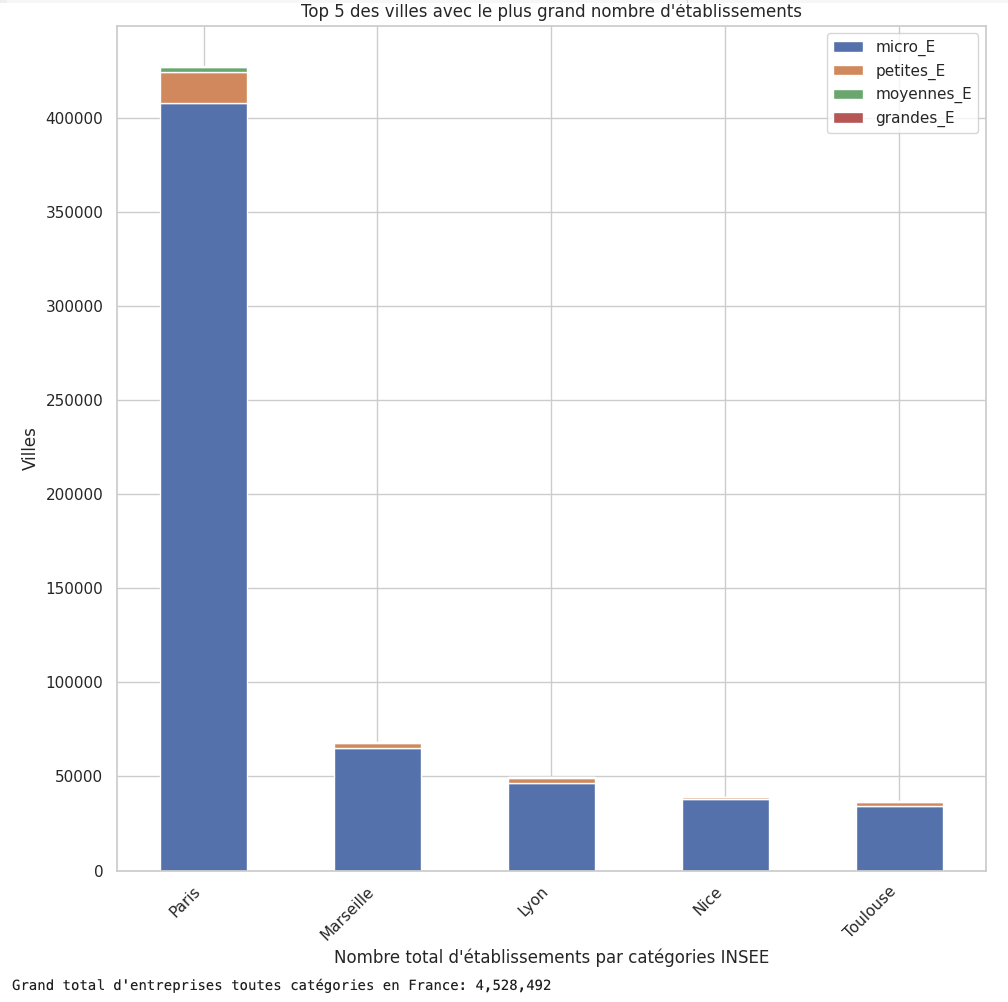
Comme sur le df\_etab\_effectifs, nous allons agréger des colonnes, puis supprimer

et / ou renommer des colonnes sur ce df

* Notre objectif est de réduire au maximum le nombre de colonnes en supprimant les colonnes qui ont des valeurs redondantes et qui risquent de fausser notre modèle de machine learning
* Nous allons enfin standardiser nos données à des fins de modélisation. Nous allons réaliser une matrice de corrélation pour préparer par la suite la partie Feature engineering, l’encodage des données, la standardisation. Ces étapes vont nous permettre de réaliser plusieurs modélisations dans le but de choisir celles qui s'adaptent parfaitement à notre dataFrame.
* Et nous serons enfin à même de déterminer une ou plusieurs valeurs cibles (nous envisageons dors et déjà les ‘salaires’)

**c)** **Data-visualisations et avis métier**

* Top 5 des villes avec le nombre d'établissements par catégorie (selon Insee)

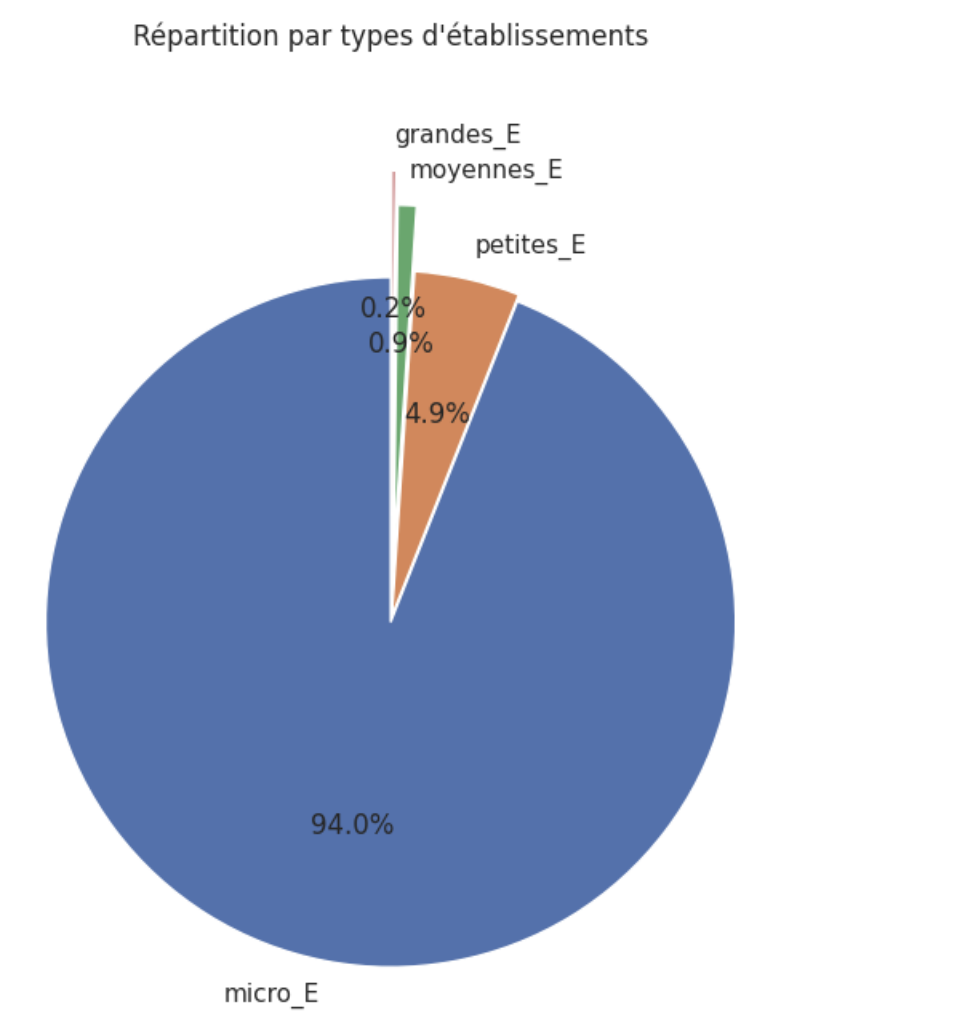


**Conclusion:**

On constate tout d'abord une prédominance du nombre de micro-entreprises dans le tissu économique français

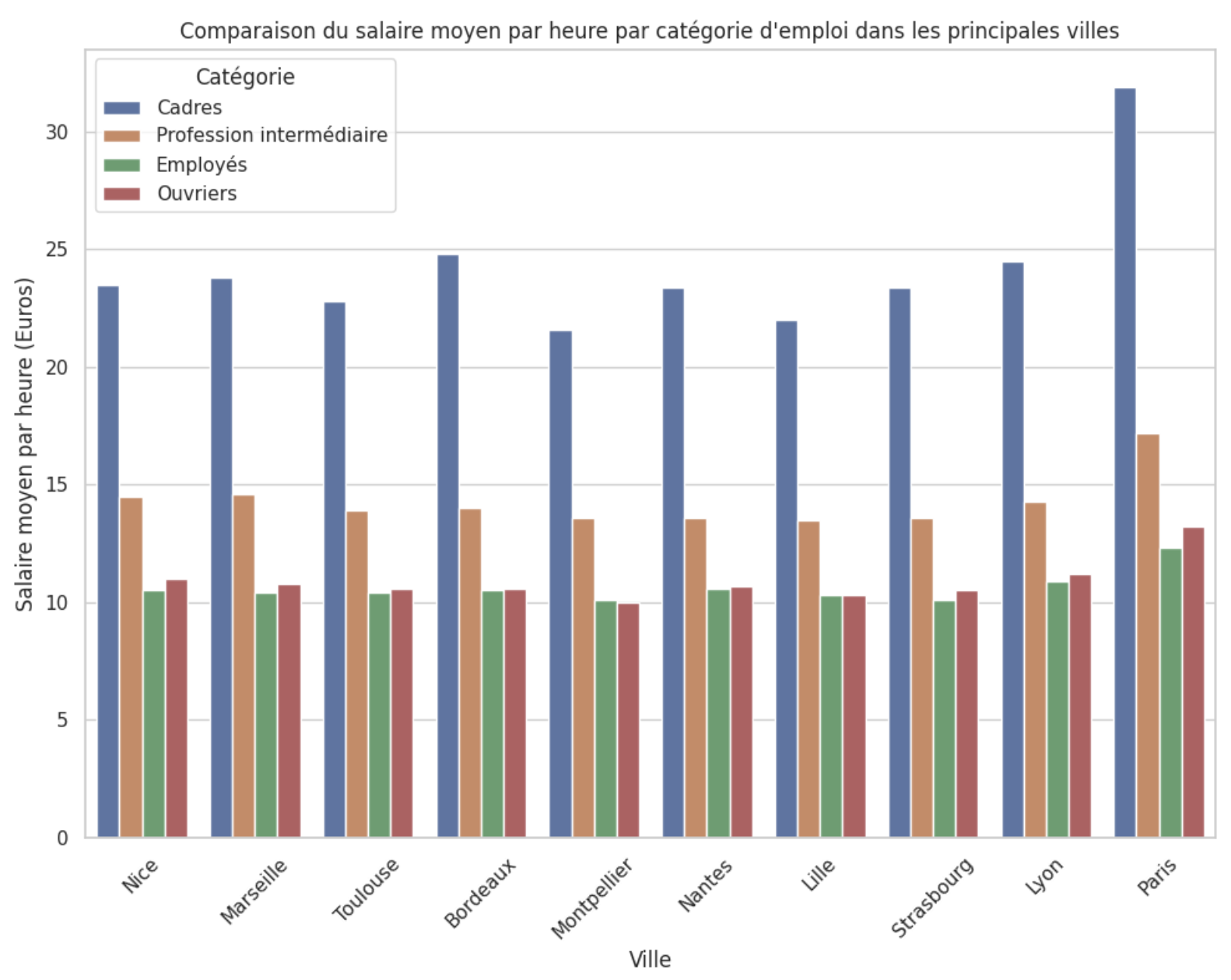
On remarque également que les entreprises sont polarisées autour des grandes villes :

* Paris représente 9% des entreprises en France (et pèse pour 69% du top 5)
* Marseille représente 2% des entreprises en France (et pèse pour 11% du top 5)
* Tandis que Lyon, Nice et Toulouse représentent chacune environ 1% des entreprises en France (et pèse pour 6% du top 5)
* Répartition par types d'établissements



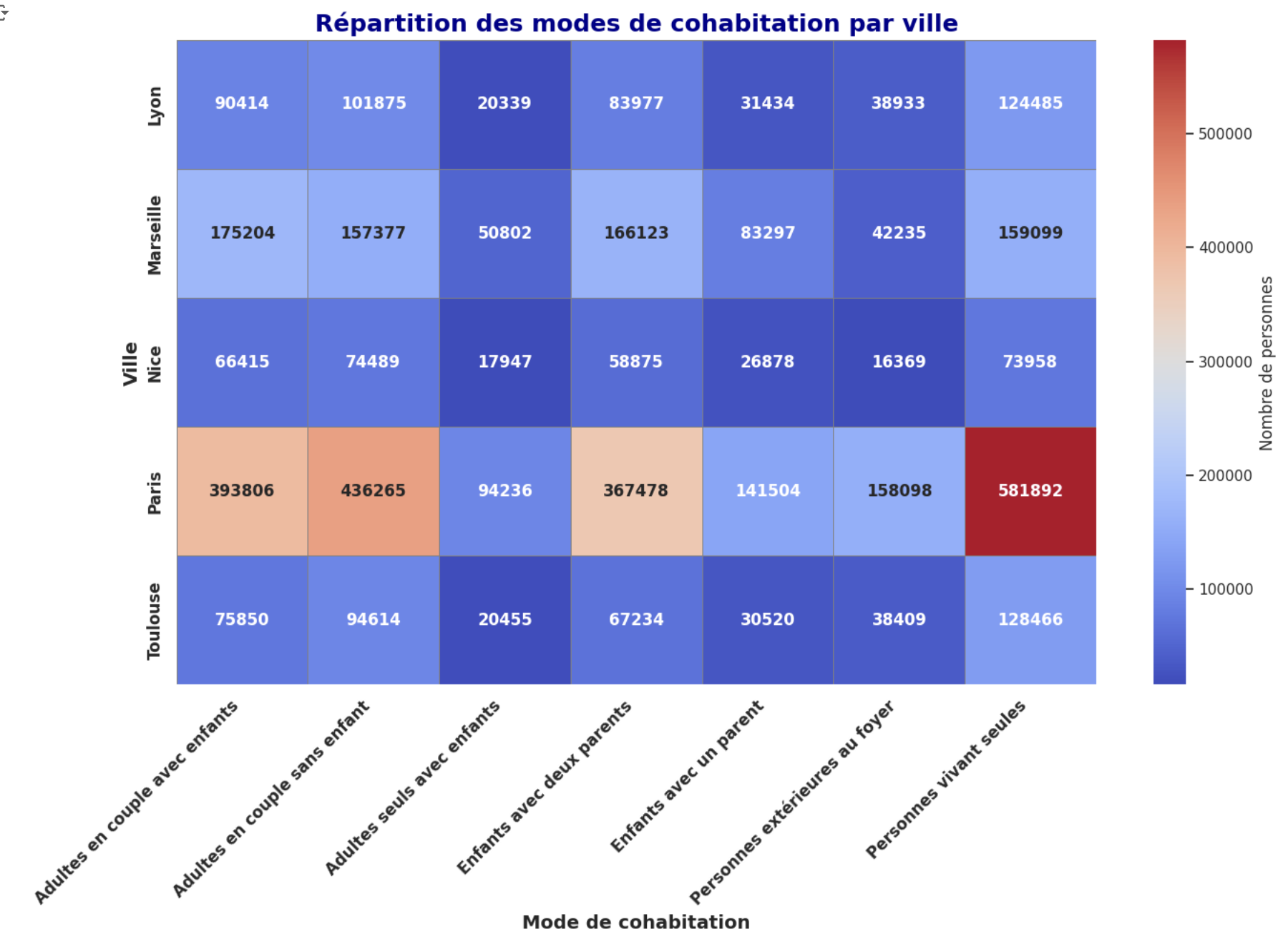
**Conclusion:**

* On remarque ici une prédominance de la micro entreprise, qui représente 94 % du nombre d'entreprises en France
* Suivit de loin par les petites entreprises, qui constituent 4,9% du nombre d'entreprises
* Les moyennes et grandes entreprises ne représentent que 1% du nombre d'entreprises en France
* **Comparaison du salaire moyen / heure / catégorie d'emploi dans les principales villes**



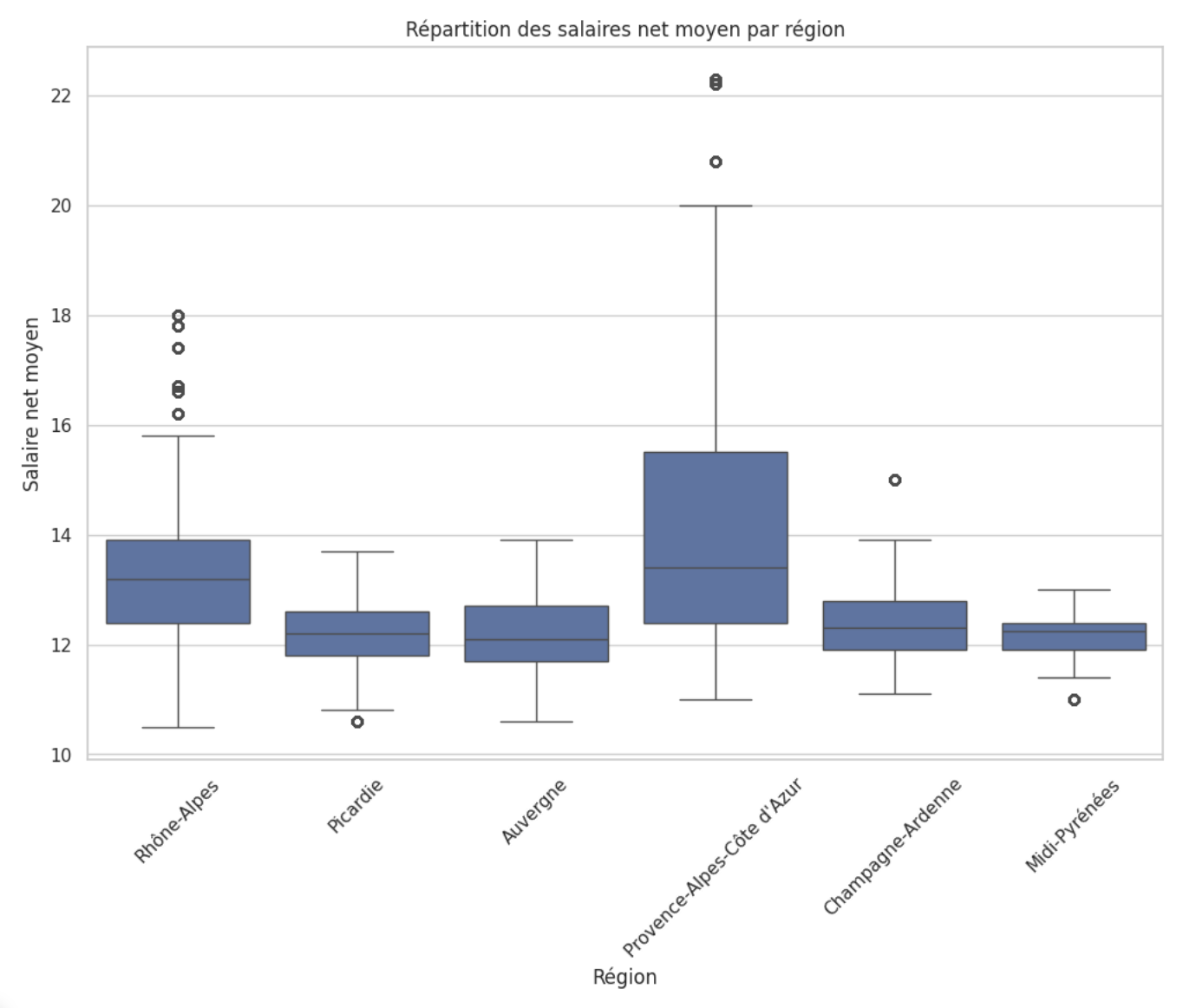
**Conclusion:**

* Sur le salaire moyen horaire des cadres : il varie entre 22 et 33 euros de l'heure
  + On constate une nette disparité entre le salaire moyen horaire des cadres et ceux des autres catégories professionnelles
  + De la même façon ce salaire moyen horaire des cadres varient fortement d'une ville comme Paris par rapport à Montpellier
* Sur le salaire moyen horaire des professions intermédiaires: il varie entre 12 et 16 euros de l'heure
  + Les écarts sur cette catégorie sont moindres
* Sur le salaire moyen horaire des employés et ouvriers : il se situe entre 10 et 14 euros de l'heure
* **Répartition des modes de cohabitation par ville**



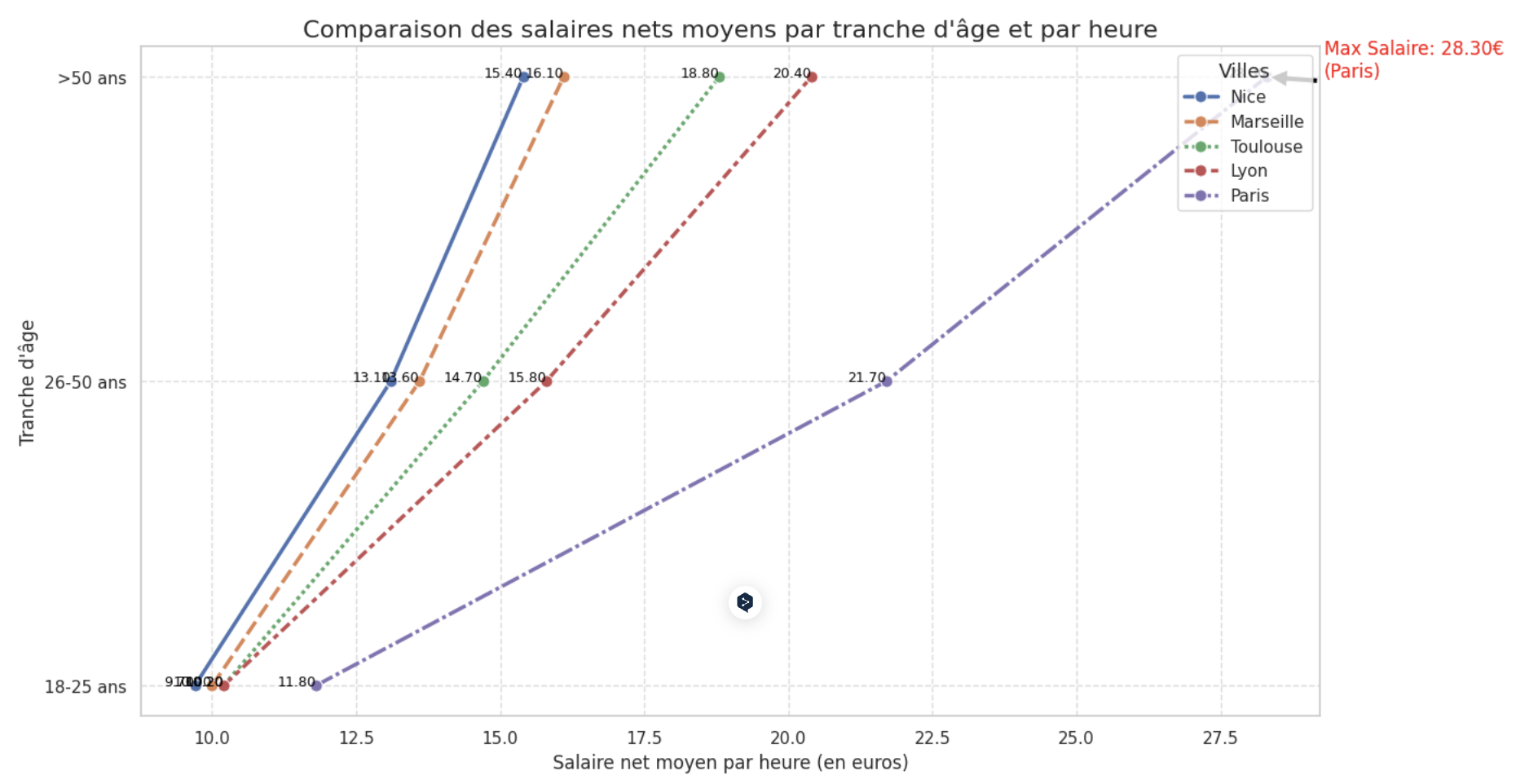
**Conclusion:**

* Paris montre une tendance marquée de "Personnes vivant seules" par rapport aux autres villes, indiquant peut-être une vie urbaine plus isolée.
* En revanche, les autres villes comme Lyon et Marseille montrent une répartition plus équilibrée entre les différents modes de cohabitation, avec une proportion significative "d'Adultes en couple" et "Enfants avec parents".
* **Box plot salaires net moyen par heure et par région**



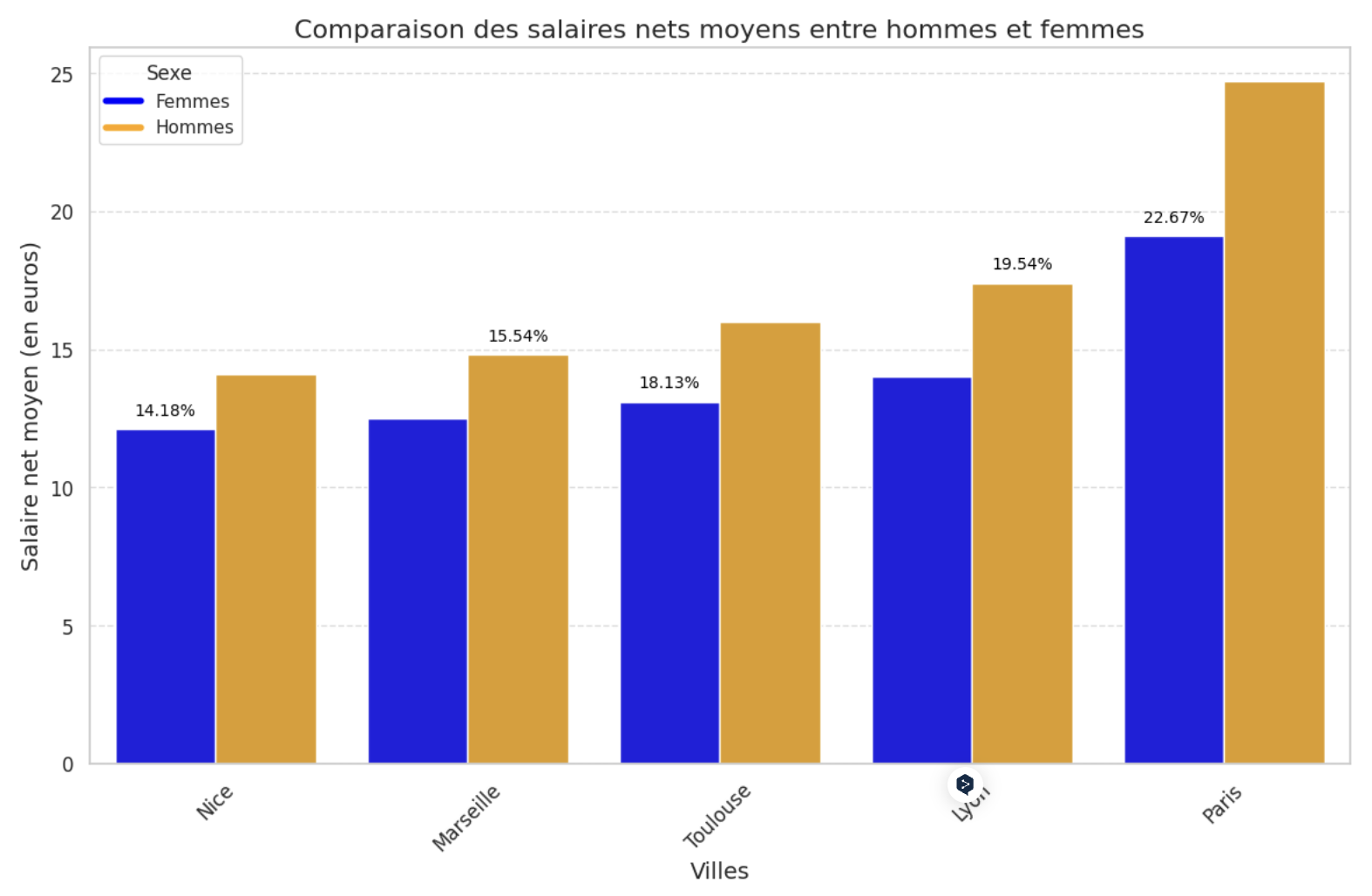
**Conclusion:**

* Le graphique montre que la région Provence-Alpes-Côte d'Azur a la plus grande dispersion des salaires nets moyens, avec une médiane plus élevée que les autres régions. Rhône-Alpes et Auvergne ont des distributions de salaires similaires, bien que Rhône-Alpes ait plus de valeurs aberrantes élevées.
* Picardie, Champagne-Ardenne et Midi-Pyrénées ont des salaires nets moyens plus bas et des dispersions plus faibles. Cela suggère que les salaires nets moyens varient considérablement selon les régions, avec certaines régions ayant des salaires plus élevés et une plus grande variabilité.
* **Lineplot comparaison salaires net moyen par heure par tranche d'âge**



**Conclusion:**

* Le graphique montre une forte corrélation entre l'âge et le salaire horaire moyen, avec des écarts notables entre les villes. Paris se distingue par des salaires nettement plus élevés dans toutes les tranches d'âge, suggérant un marché du travail plus compétitif dans la capitale.
* En revanche, des villes comme Lyon, Toulouse, et Marseille ont des niveaux de salaires moyens plus faibles, mais cohérents avec l'évolution observée en fonction de l'âge.
* La différence salariale la plus prononcée apparaît pour les personnes de plus de 50 ans, ce qui souligne l'impact de l'expérience et de l'ancienneté sur la rémunération, particulièrement dans les grandes métropoles comme Paris.
* **Comparaison des salaires net moyens par heure entre femmes et hommes**



**Conclusion:**

* Le graphique met en évidence une discrimination salariale claire entre les sexes dans toutes les villes, avec un écart qui atteint jusqu'à 32% à Paris.
* Ces résultats montrent que malgré les différences dans les niveaux de salaire entre les villes, l'inégalité salariale entre hommes et femmes reste un problème majeur, particulièrement dans les grandes métropoles où les salaires sont plus élevés.

**d) Constats sur ces Data-visualisations**

L’analyse de ces données sur les entreprises, les salaires et les modes de vie en France met en lumière plusieurs tendances significatives :

* **Concentration géographique** : les entreprises, en particulier les micro-entreprises, sont fortement concentrées dans les grandes villes, notamment Paris. Cette polarisation urbaine a un impact direct sur les salaires et le mode de vie, avec des écarts notables entre les grandes métropoles et les autres régions.
* **Structure du tissu économique**: Le tissu économique français est dominé par les micro-entreprises, ce qui souligne la prédominance d’un modèle économique fondé sur les petites structures. Les moyennes et grandes entreprises, bien que moins nombreuses, jouent un rôle crucial dans l’économie nationale.
* **Écarts de salaires** : Les salaires varient considérablement selon la ville, la catégorie professionnelle, l’âge et le sexe. Les cadres bénéficient de rémunérations nettement supérieures aux autres catégories, et les écarts de salaires sont particulièrement marqués à Paris. Les femmes subissent une discrimination salariale significative dans toutes les villes, avec un écart salariale plus important dans les grandes métropoles.
* **Modes de vie**: les modes de vie sont différents selon les tailles des villes. Paris se distingue par une proportion élevée de personnes vivant seules, tandis que les autres villes présentent une répartition plus équilibrée des modes de cohabitation.
* **Impact de l’âge** : l’âge est un facteur déterminant dans la rémunération, avec des salaires qui augmentent généralement avec l’expérience professionnelle. Cependant, les écarts de salaires entre les différentes tranches d’âge sont plus marqués dans les grandes villes.

**En synthèse**, ces visualisations révèlent un paysage économique et social français marqué par des disparités territoriales, des inégalités salariales et une concentration des activités économiques dans les grandes villes. Les politiques publiques devraient prendre en compte ces disparités pour favoriser un développement économique plus équilibré et réduire les inégalités sociales.

**Idées d’actions futurs en attendant la mise en place d’un ou plusieurs modèles de ML:**

* Politique régionale: Il est nécessaire de mettre en place des politiques régionales plus ambitieuses pour réduire les disparités entre les territoires et favoriser le développement économique dans les régions moins dynamiques.
* Politique de l’emploi: Des mesures spécifiques doivent être prises pour lutter contre la discrimination salariales et favoriser l’égalité professionnelles entre les femmes et les hommes.
* Soutien aux entreprises: Un soutien accru aux petites et moyennes entreprises, notamment dans les régions moins favorisées, pourrait contribuer à diversifier le tissu économique et créer de l’emploi.
* Amélioration des conditions de vie: Des politiques visant à améliorer les conditions de vie dans les grandes villes, notamment en matière de logement et de transport, pourraient contribuer à réduire les tensions sociales et à favoriser un développement urbain plus équilibré.

**e)** **Tests et avis métiers**

Dans le cadre de notre projet, nous avons mis en place une série de tests de corrélation et d'analyses statistiques afin d’explorer les relations entre différentes variables quantitatives et d’identifier les facteurs les plus influents sur les inégalités et les salaires en France.

* ***Test de Pearson entre deux variables quantitatives avec une corrélation entre le salaire net moyen et les grandes entreprises***

Nous cherchons une corrélation entre le salaire net moyen et la taille des entreprises (grandes et moyennes entreprises)

Nous obtenons comme résultats :

* Corrélation de Pearson : 0.10153681255648732
* Valeur p : 3.0130709110800824e-13

**Conclusion:**

On rejette la première hypothèse et on garde H1 car le résultat indique une corrélation très faible de 0,10 entre le salaire médian et le nombre d'entreprises de grande taille, mais cette relation est statistiquement significative avec une valeur p extrêmement faible (3.01)..

* ***Test de Pearson entre deux variables quantitative avec une corrélation entre le salaire net moyen et les moyennes entreprises***

Nous cherchons une corrélation entre le salaire net moyen et les moyennes entreprises.

Nous obtenons comme résultats :

* Corrélation de Pearson : 0.07669166894070512
* Valeur p : 3.734997811020279e-08

**Conclusion:**

La corrélation de Pearson entre le salaire net moyen (SNHM14) et le nombre de moyennes entreprises (moyennes\_E) est extrêmement faible (0,07), indiquant une relation pratiquement inexistante entre ces variables.

Cependant, la valeur p (3.7) est bien en dessous du seuil de 0,05 ce qui signifie que cette faible corrélation est statistiquement significative

* ***Test de Pearson entre deux variables quantitative avec une corrélation entre le salaire net moyen par heure pour un employé et le nombre de personnes non scolarisées de 15 ans ou plus titulaires d'un diplôme Bac en 2021***

Nous cherchons une corrélation entre le salaire net moyen et les facteurs démographiques, et le niveau d'éducation.

Nous obtenons comme résultats :

* Corrélation de Pearson : 0.03607033501880219
* Valeur p : 0.009731554620658845

**Conclusion:**

La corrélation de Pearson obtenue (0.03) indique une relation extrêmement faible entre le salaire net moyen par heure pour un employé et le nombre de personnes non scolarisées de 15 ans ou plus titulaires d'un diplôme Bac.

Bien que la valeur p très faible (0.01) signifie que cette relation est statistiquement significative, la force de cette corrélation est si faible qu'elle n'a probablement pas de pertinence pratique.

* ***Test de spearman entre les variables entreprise taille micro\_E et le salaire net moyen***

Nous cherchons une corrélation entre le salaire net moyen et la micro-entreprise

Nous obtenons comme résultats :

* Coefficient de corrélation de Spearman : 0.10841591834550372
* Valeur p : 6.6583665793122616e-15

**Conclusion:**

La corrélation de Spearman entre le salaire médian et la variable taille micro\_E est de 0,10, ce qui indique une faible corrélation positive.

Cependant, la valeur p extrêmement faible (6.65) indique que cette corrélation est statistiquement significative.

Cela signifie qu'il y a effectivement une relation entre les deux variables, bien que cette relation soit faible

* **Test anova entre les régions du territoire**

|  | **sum\_sq** | **df** | **F** | **PR(>F)** |
| --- | --- | --- | --- | --- |
| C(REG) | 9661.329553 | 25.0 | 82.37238 | 0.0 |
| Residual | 23973.760977 | 5110.0 | NaN | NaN |

**Conclusion:**

Ces résultats ANOVA montrent que le test est extrêmement significatif. Le F-score élevé (82) et la p-value de 0 indiquent qu'il y a une influence significative des régions (REG) sur le salaire net moyen (SNHM14).

Ainsi, nous rejetons l'hypothèse nulle (H0) et acceptons l'hypothèse alternative (H1).

**f) Constats sur ces tests**

Les analyses et tests que nous avons menés nous indiquent des corrélations existantes mais faibles.

Il convient donc de pousser plus en avant nos analyses et recherches, afin de définir clairement une ou plusieurs variables cibles pour notre modèle de ML

**g)** **Conclusion sur ce rapport d’exploration**

Nous avons donc :

* exploré et analysé les 4 jeux de données,
* enrichi notre jeu de données,
* nettoyé et pré-processé les dataframes,
* créé et analysé un premier set de data visualisation,
* procédé à des tests afin d’établir des corrélations, et déterminer des valeurs cibles.

**Nos visualisations révèlent un paysage économique et social français marqué par des disparités territoriales, des inégalités salariales et une concentration des activités économiques dans les grandes villes.**

**03) Rapport de modélisation**

1. **Première itération de la matrice de corrélation**

Après analyses de ces premiers tests et graphiques, nous allons maintenant chercher à prédire notre valeur cible, en démarrant tout d’abord par la phase de pré-processing.

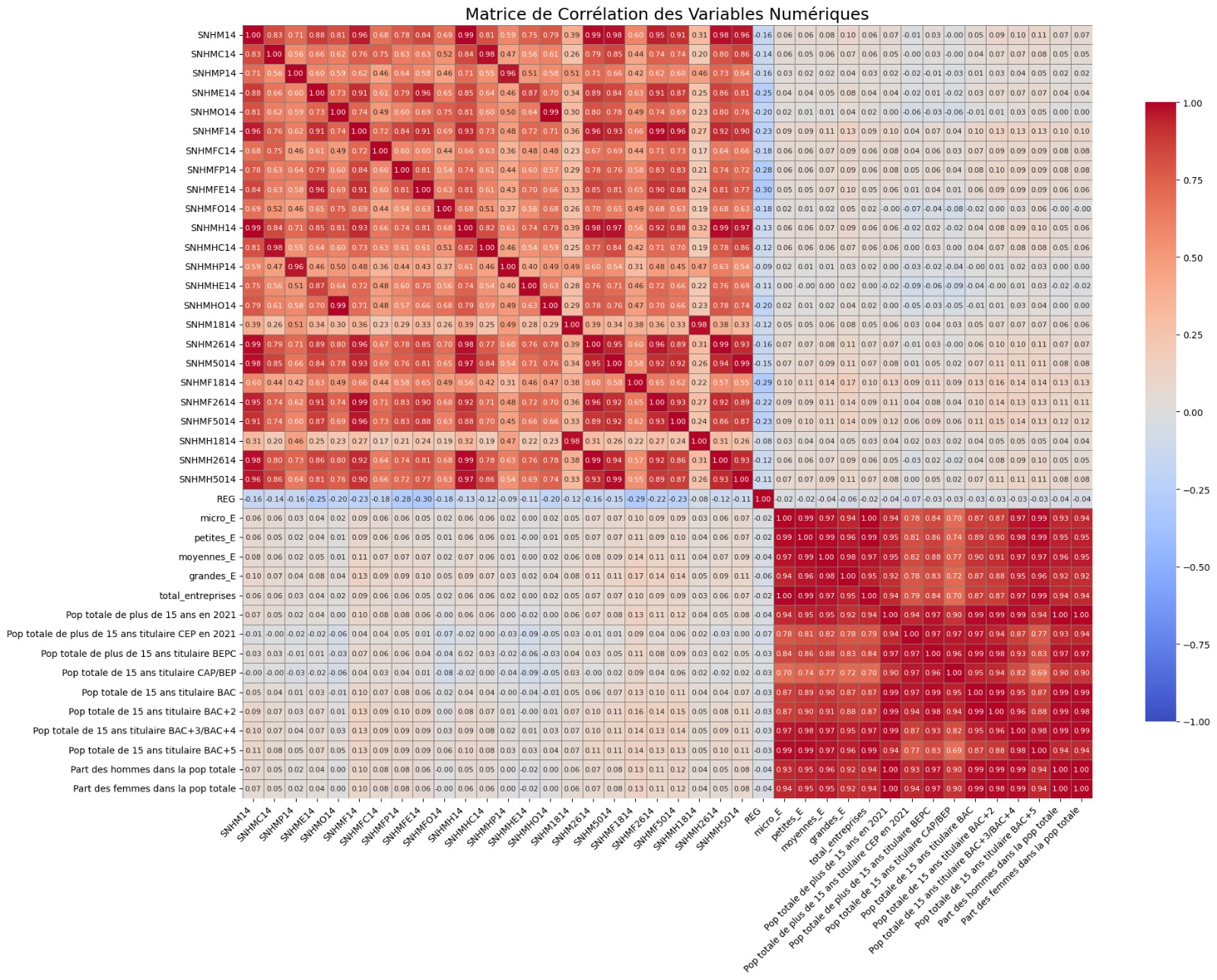
Pour cela, nous allons créer successivement plusieurs matrices de corrélation, afin d’analyser les relations entre les variables de notre jeu de données

Nous allons les utiliser afin de:

* Détecter des relations linéaires et identifier les variables qui ont une corrélation élevée, positive ou négative (ce qui peut indiquer qu'elles partagent une certaine redondance en termes d'information)
* Réduire la colinéarité (des variables très corrélées augmentent la complexité du modèle sans apporter de nouvelles informations)
* Choisir les variables les plus pertinentes (si certaines variables sont faiblement corrélées avec la variable cible mais fortement corrélées entre elles, il peut être intéressant d'en sélectionner une seule ou d'envisager une transformation)
* Comprendre la relation dans les données (si deux variables sont corrélées, il peut être pertinent de créer une nouvelle variable qui les combine pour capturer leur interaction de manière plus explicite)

Nous allons donc nous servir de ces matrices pour mieux comprendre les relations entre les variables, pour simplifier le modèle et améliorer ses performances

Voici la première matrice de corrélation:



**Conclusion:**

Après lecture de cette première matrice, nous décidons des actions suivantes:

* Afin d’éviter une colinéarité entre les variables relatives aux diplômes, nous décidons de supprimer les variables suivantes:
  + Pop totale de plus de 15 ans titulaire CEP en 2021
  + Pop totale de plus de 15 ans titulaire BEPC
  + Pop totale de 15 ans titulaire BAC+3/BAC+4
  + Pop totale de 15 ans titulaire BAC+2
  + Part des hommes dans la pop totale
  + Part des femmes dans la pop totale

Et de ne garder que:

* + Pop Totale de + 15 ans en 2021
  + Pop Totale de + 15 ans en 2021 titulaires CAP et BEP
  + Pop Totale de + 15 ans en 2021 titulaire BAC
  + Pop Totale de + 15 ans en 2021 titulaire BAC +5
* Pour éviter une colinéarité entre les variables relatives aux salaires, nous décidons de supprimer les variables suivantes:

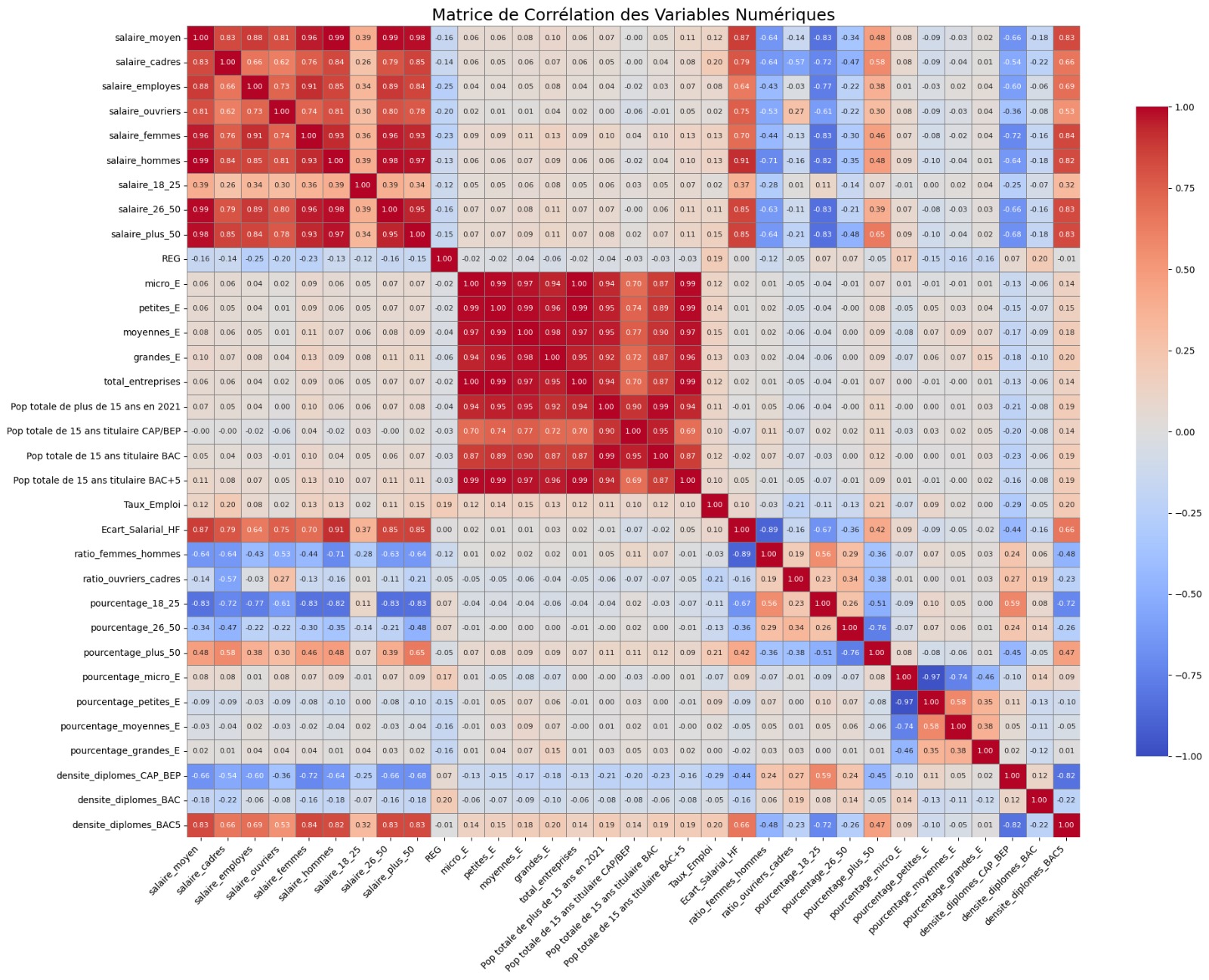
SNHMP14, SNHMFC14, SNHMFP14, SNHMFE14, SNHMFO14, SNHMHC14, SNHMHP14, SNHMHE14, SNHMHO14, SNHMF1814, SNHMF2614, SNHMF5014, SNHMH1814, SNHMH2614, SNHMH5014

En vue de la seconde matrice de corrélation:

* Nous allons créer dans notre dataframe de nouvelles variables pour alimenter notre futur modèle de Machine Learning:
  + Taux emploi = total\_entreprises / Pop totale de plus de 15 ans en 2021\*100
  + Ecart\_Salarial\_HF = salaire\_hommes - salaire\_femmes
* Nous allons également calculer des ratios:
  + ratio\_femmes\_hommes = salaire\_femmes / salaire\_hommes
  + ratio\_ouvriers\_cadres = salaire\_ouvriers / salaire\_cadres
* Puis nous allons calculer le pourcentage de chaque groupe démographique:
  + pourcentage\_18\_25 = salaire\_18\_25 / salaire\_moyen
  + pourcentage\_26\_50 = salaire\_26\_50 / salaire\_moyen
  + pourcentage\_plus\_50 = salaire\_plus\_50 / salaire\_moyen
* Puis calculer le pourcentage des types d'entreprises:
  + pourcentage\_micro\_E = micro\_E / total\_entreprises
  + pourcentage\_petites\_E = petites\_E / total\_entreprises
  + pourcentage\_moyennes\_E = moyennes\_E / total\_entreprises
  + pourcentage\_grandes\_E = grandes\_E / total\_entreprises
* Et enfin calculer la densité de diplômés
  + densite\_diplomes\_CAP\_BEP = Pop totale de 15 ans titulaire CAP/BEP / Pop totale de plus de 15 ans en 2021
  + densite\_diplomes\_BAC = Pop totale de 15 ans titulaire BAC / Pop totale de plus de 15 ans en 2021
  + densite\_diplomes\_BAC5 = Pop totale de 15 ans titulaire BAC+5 / Pop totale de plus de 15 ans en 2021

1. **Seconde itération de la matrice de corrélation**

Voici la seconde matrice de corrélation:



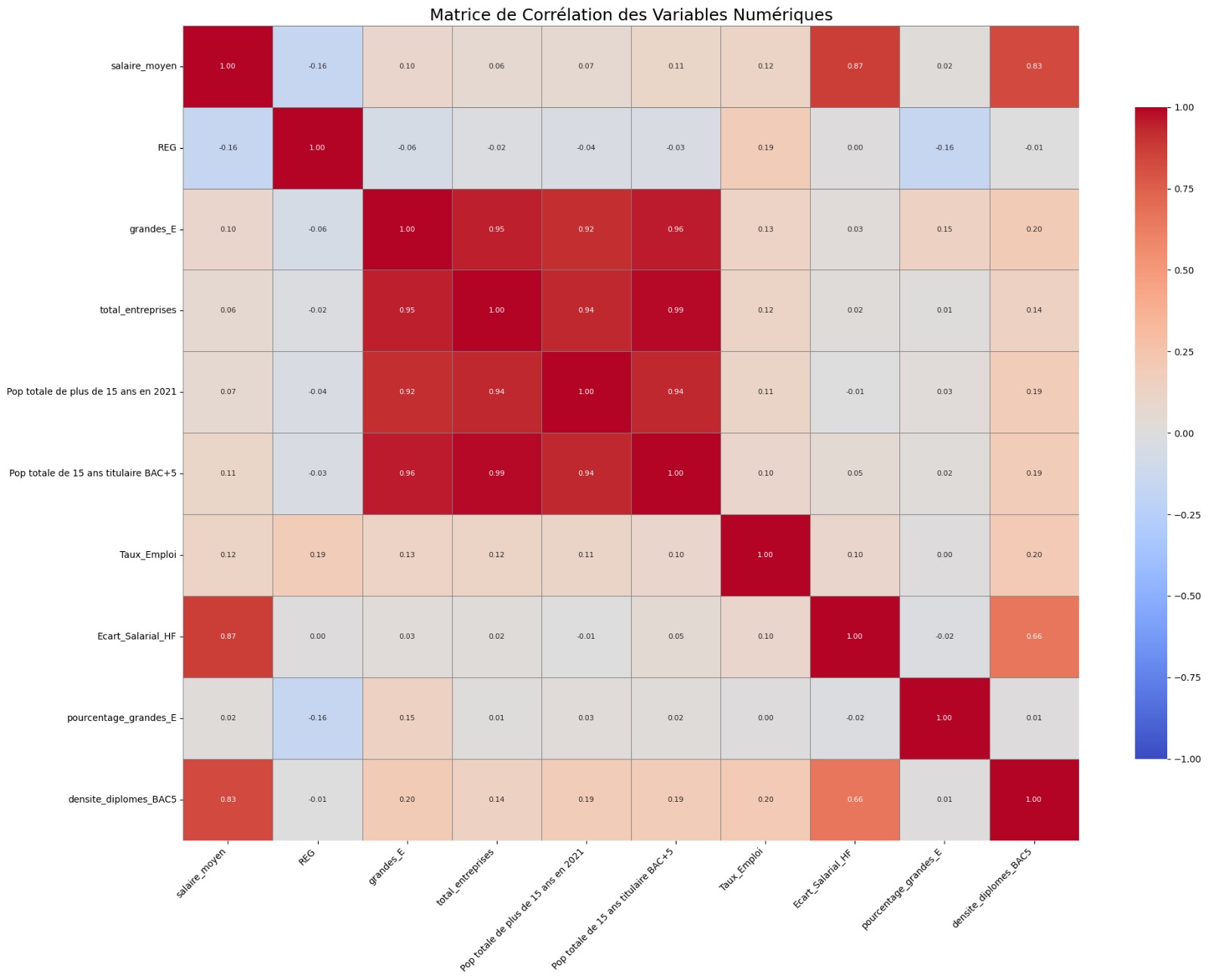
**Conclusion:**

Après lecture de cette seconde matrice de corrélation, et en vue de la troisième et dernière, nous décidons des actions suivantes:

* Pour éviter une colinéarité avec le salaire moyen, que nous définissons comme notre variable cible, nous décidons de supprimer les variables suivantes:
  + salaire\_cadres,
  + salaire\_employes,
  + salaire\_ouvriers,
  + salaire\_18\_25,
  + salaire\_26\_50,
  + salaire\_femmes,
  + salaire\_hommes,
  + ratio\_femmes\_hommes,
  + pourcentage\_plus\_50
  + salaire\_plus\_50
* Pour éviter une colinéarité entre les variables relatives aux salaires, nous décidons de supprimer les variables suivantes relatives à la taille des entreprises et aux pourcentages pour ne garder que Total Entreprises et Grandes Entreprises:
  + micro\_E
  + petites\_E
  + moyennes\_E
  + pourcentage\_18\_25
  + pourcentage\_26\_50
  + pourcentage\_micro\_E
  + pourcentage\_petites\_E
  + pourcentage\_moyennes
* Toujours pour des problèmes de colinéarité, nous supprimons les variables relatives à la population totale et aux diplômes:
  + Pop totale de 15 ans titulaire CAP/BEP
  + Pop totale de 15 ans titulaire BAC
  + ratio\_ouvriers\_cadres
  + densite\_diplomes\_CAP\_BEP
  + densite\_diplomes\_BAC

1. **Troisième itération de la matrice de corrélation**

Voici la troisième et dernière matrice de corrélation: nous choisissons ces variables pour alimenter notre modèle de Machine Learning:



1. **Machine Learning et entraînement des modèles**

Entraînons et testons le Machine Learning selon les quatres modèles suivants.

Calculons et concentrons-nous sur les résultats du R-squared (R²)

| **Modèles** | Résultats du Training | Résultats du Test |
| --- | --- | --- |
| Régression Linéaire | 1.0000000000000000 | 0.7289496637168904 |
| XG Boost | 0.9734546606212742 | 0.9135809281163670 |
| Random Forest | 0.9903104229610412 | 0.9117021770765382 |
| K-Nearest Neighbors | 0.9038736473331437 | 0.8706452554901113 |

**Conclusion**

* Le modèle Random Forest semble être le choix le plus robuste pour cet entraînement de l'ensemble de données, avec un excellent R² et des erreurs plus faibles par rapport à XG Boost et KNN.
* La régression linéaire a un score parfait, mais cela pourrait nécessiter une validation croisée plus approfondie pour vérifier sa généralisation.

Afin d’affiner nos prévisions, nous allons appliquer les actions suivantes, puis re-calculer le R²:

* **Vérification de la variabilité des données,** afin de s'assurer que les données d'entraînement et de test sont bien équilibrées et qu'il n'y a pas trop de similarités ou de fuites entre ces ensembles.
* **Réduction de la complexité des modèles**
* **Contrôle du sur-apprentissage:**
  + On va ajouter de la régularisation pour les modèles complexes (comme XG Boost ou Random Forest).
  + Nous allons également ajuster les hyperparamètres (comme le nombre d'arbres pour Random Forest, ou le nombre de voisins pour KNN)
  + Nous allons aussi augmenter la taille du test set
* **Cross-validation** pour mieux évaluer la performance des modèles sur plusieurs sous-ensembles de vos données, et voir si ces performances élevées persistent

La méthode de Cross-validation nous donne les résultats suivants:

* R² moyen (**régression linéaire**) avec validation croisée : 0.6350068233380128
* R² moyen (**XG Boost**) avec validation croisée : 0.862061830705661
* R² moyen (**Random Forest**) avec validation croisée : 0.8417870493429259
* R² moyen (**KNN**) avec validation croisée : 0.7696211735144691

**Conclusion:**

* Nous pouvons affirmer que le modèle XG Boost se distingue comme le meilleur choix pour ce jeu de données, suivi de près par Random Forest.
* Nous constatons que les modèles de régression linéaire et KNN ont des performances correctes, mais sont inférieures, probablement en raison de leur limitation à modéliser des relations non linéaires et des effets d'interaction entre les variables.
* Ces métriques montrent que votre modèle **XG Boost** a une capacité de prédiction très précise.

1. **Choix du modèle final**

Évaluons maintenant les performances finales de XG BOOST sur l'ensemble du test:

* **Mean Squared Error:** 0.4100964477836892
* **Root Mean Squared Error**: 0.6403877323806955
* **Mean Absolute Error**: 0.47074711890595056
* **R-squared**: 0.9195168530807631

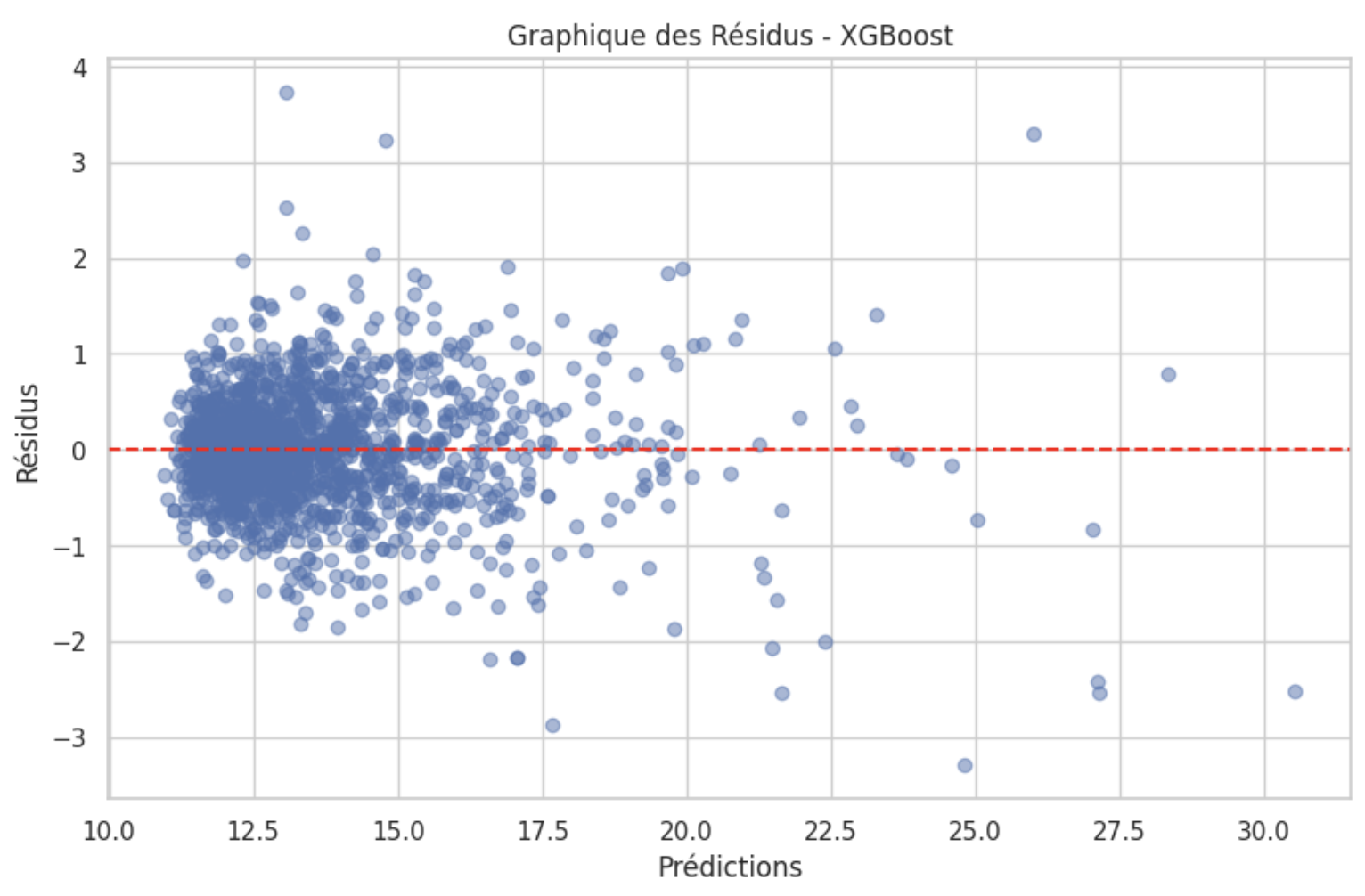
**Conclusion:**

Ces résultats nous permettent de valider le choix de XG BOOST comme modèle de Machine Learning

1. **Analyse des graphiques des erreurs, de répartition, et des résidus**

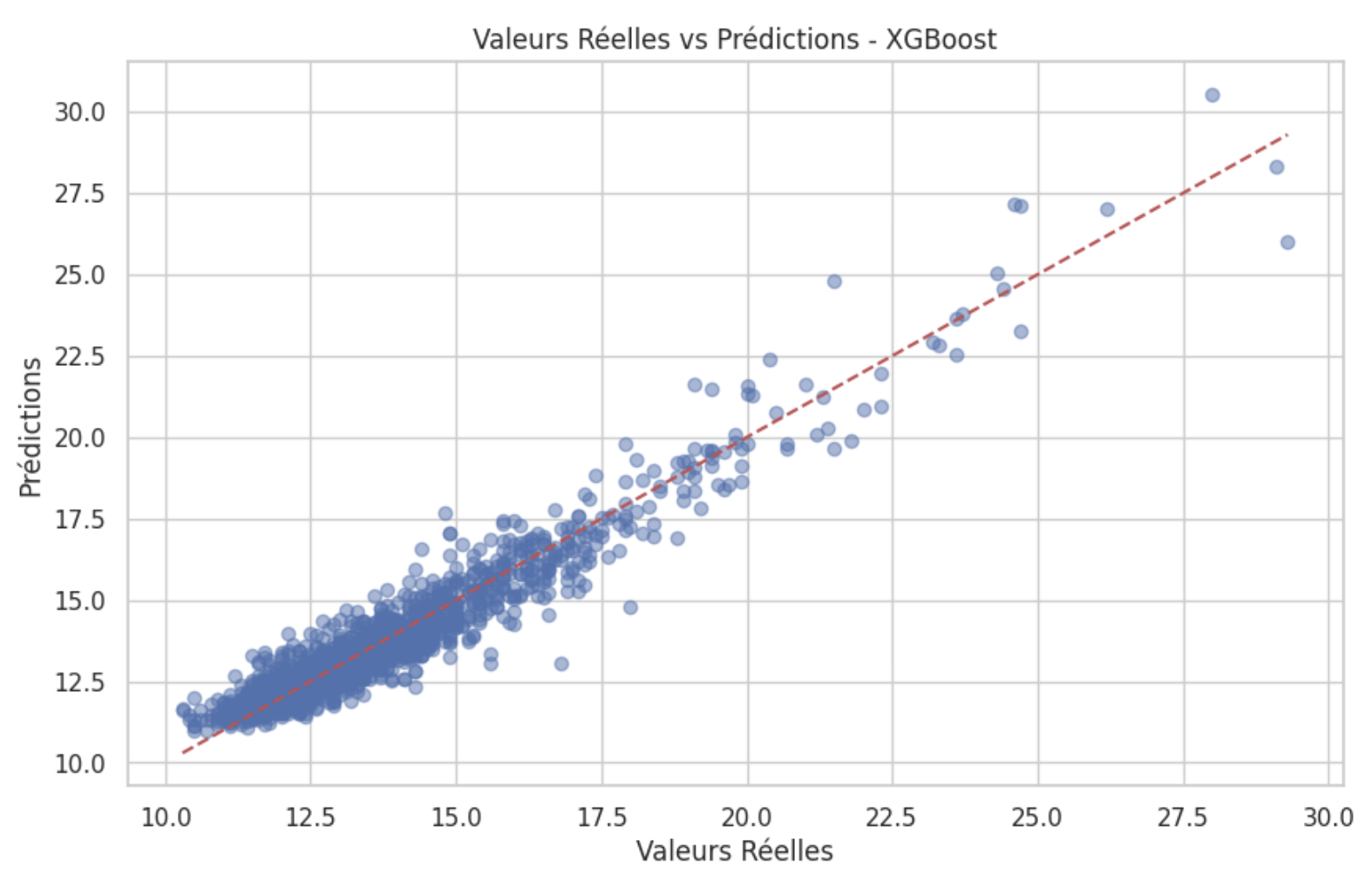
Etudions maintenant les graphiques suivants afin de confirmer définitivement notre choix:

* **Graphique des Erreurs de Prédiction (Résidus)**



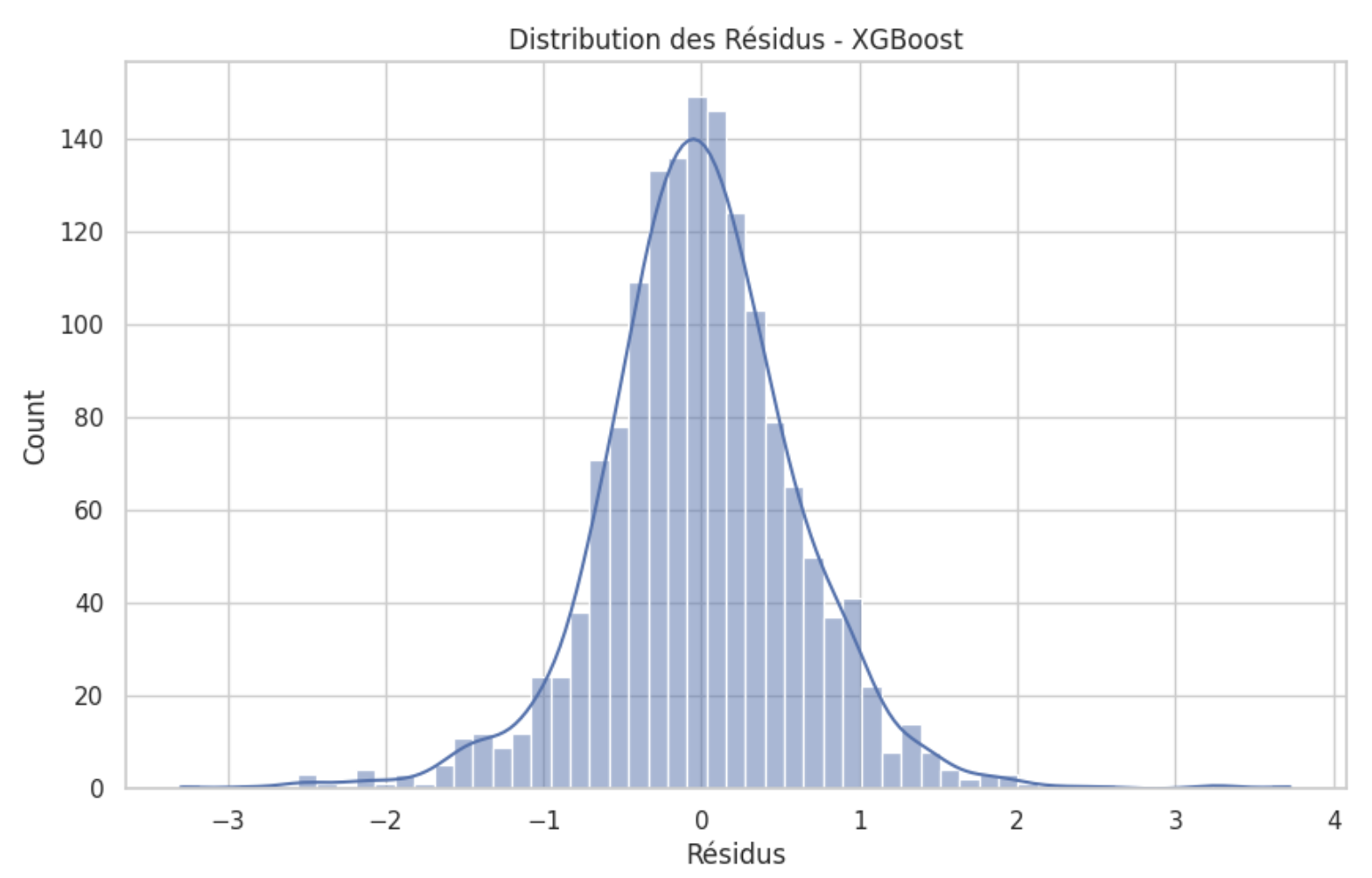
**Conclusion:**

* Le graphique des résidus présenté montre une dispersion aléatoire des points autour de la ligne horizontale à zéro, ce qui est un bon indicateur.
* Cela suggère que le modèle XGBoost capture bien les tendances sous-jacentes des données et que les erreurs de prédiction sont distribuées de manière aléatoire.
* **Graphique de Répartition des Valeurs Réelles vs Prédictions**



**Conclusion:**

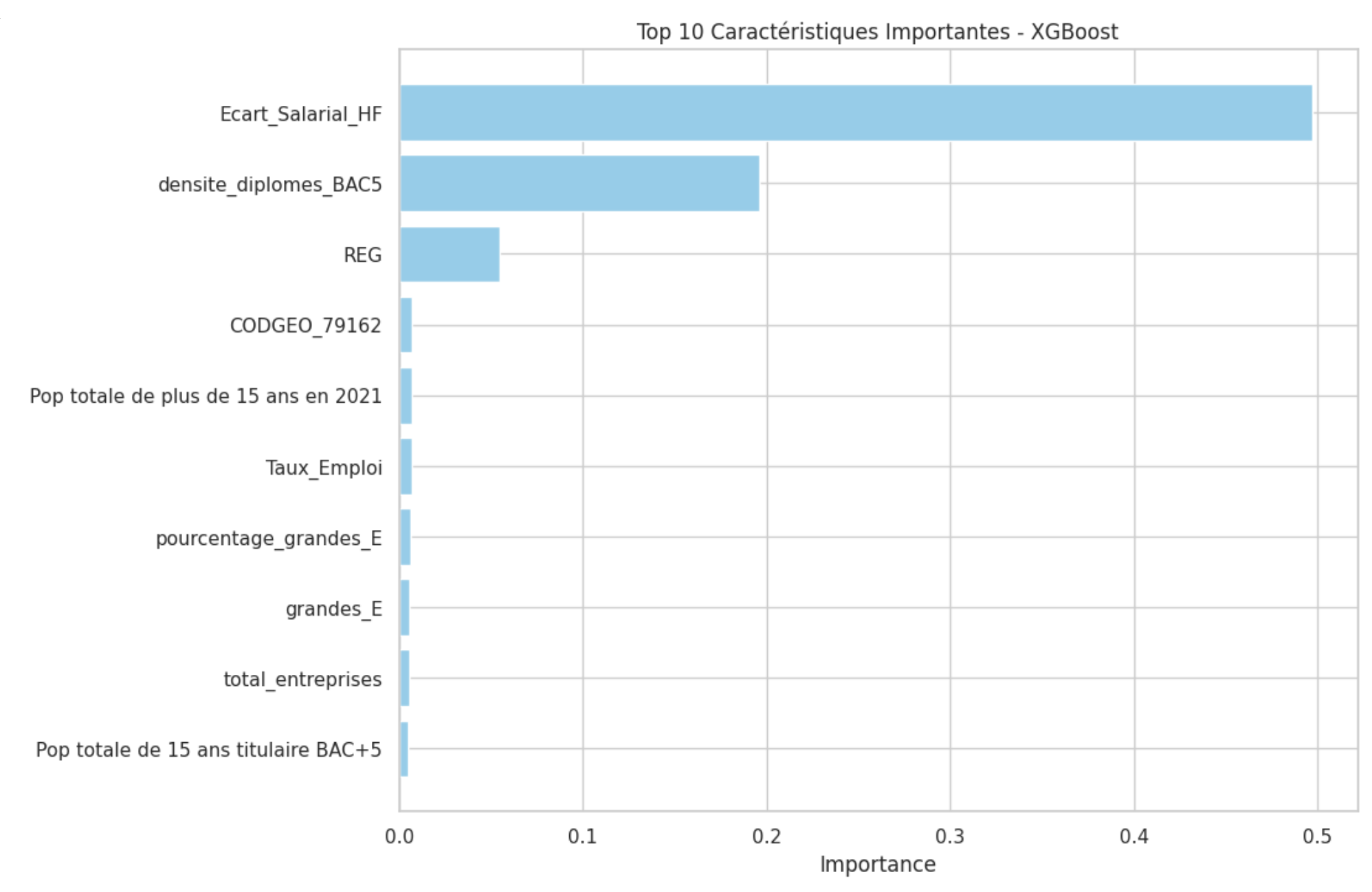
* Notre modèle semble globalement bien calibré, car la majorité des résidus sont proches de zéro.
* Cela indique que le modèle ne fait pas d'erreurs systématiques et capture bien la relation entre les caractéristiques et notre variable cible (salaire moyen).
* **Graphique de distribution des Résidus**

****

**Conclusion:**

* Ce graphique de distribution des résidus suggère que le modèle XGBoost offre une bonne performance sur l'ensemble des données.
* La distribution des erreurs est relativement proche de la normale, ce qui est un indicateur de qualité pour un modèle de régression

1. **Graphique d’interprétabilité**



Nous constatons que les 3 variables Ecart\_Salarial\_HF, Densité\_diplomesBAC\_5 et REG sont les variables clefs qui vont le plus influencer notre modèle de Machine Learning

1. **Conclusion scientifique et métier**

* Dans notre analyse, l’écart salarial entre hommes et femmes se révèle être le facteur le plus influent dans la prédiction des salaires moyens, confirmant qu’il s’agit d’un déterminant majeur des inégalités salariales.

La densité de diplômés Bac+5 apparaît également comme un autre facteur critique, ce qui souligne l’importance du niveau d’éducation dans la détermination des salaires

Les autres caractéristiques, telles que le taux d'emploi, le nombre total d'entreprises ou les codes géographiques, ont un impact moins important mais contribuent néanmoins à la précision du modèle.

* Notre modèle XGBoost accorde une importance prépondérante à l'écart salarial et au niveau d'éducation pour effectuer ses prédictions, qui constituent des leviers importants pour agir sur les inégalités
* D’un point de vue métier, cela signifie que les politiques visant à réduire les écarts salariaux devraient se concentrer sur l’égalité des genres dans les salaires et mettre l’accent sur des politiques d’éducation et de formation.
* En conséquence, il serait pertinent de recommander aux entreprises et aux collectivités de développer des initiatives pour améliorer l'accès à la formation supérieure et d'encourager des pratiques de rémunération équitable.

Nous pourrions recommander des programmes de formation spécifiques pour les femmes dans les secteurs à forte croissance, qui pourraient à la fois réduire l'écart salarial et favoriser une montée en compétences dans des domaines bien rémunérés.