

☼ SUNBURST ☼

**S2.06 Analyse de données, reporting et datavisualisation**

RAPPORT D’ANALYSE

**Présentation des données et du nettoyage éventuel**

30/04/2024 **Groupe : Galton**

**Thevasenan Satujan Berkat Arselen Avenel Kyllian Tep Mathieu**

**Diallo Thierno**

[**Résumé 3**](#_qg85d52qe279)

[**Summary 3**](#_qx7v1ea039v0)

[**I. Introduction 4**](#_44jx72mnl5z)

[**II. Présentation des thèmes 4**](#_zdy8qk60smrn)

[1. Évolution et structure de la population 4](#_a9fkv4o4e3k0)

[2. Caractéristique de l’emploi au sens du recensement 5](#_vzrycorq2kuf)

[3. Couples - Familles - Ménages 5](#_lq686kffvugn)

[4. Évolution et structure de l’entreprise 6](#_td7l808pk0ic)

[5. Diplôme et Formation 6](#_kn89pkx8thx0)

[6. Caractéristiques des établissements 7](#_ayrcytdk0rvj)

[7. Évolution des revenus et pauvreté des ménages 7](#_enyiko4i3ago)

[8. Naissances et décès domiciliés 8](#_n5bdfjpcd4ao)

[9. Population active, emploi et chômage au sens du recensement 8](#_4wnq5t2v9kss)

[10. Salaire et revenus d’activité 9](#_ludtdnu8xu9s)

[11. Tourisme 9](#_muc5gywbxymw)

[12. Logement 10](#_hdbpltfi74my)

[**III. Nettoyage éventuel 10**](#_xp947clajw4n)

[**IV. Conclusion 11**](#_aivur5hsgsra)

[**Bibliographie 12**](#_cdivraksdse3)

# Résumé

Dans ce rapport nous décrivons d’une part la base données issus de l’insee et d’autre part du nettoyage éventuel à effectuer

Ce document présente plusieurs thèmes socio-économiques allant de la démographie à l'emploi, en passant par les entreprises, les revenus et le logement. Chaque thème est abordé à travers une série de variables décrivant différents aspects comme la population par âge et genre des communes de France.

# Summary

In this report, we describe the insee database and the cleaning work that needs to be done.

This document presents several socio-economic themes, ranging from demographics to employment, business, income and housing. Each theme is approached through a series of variables describing different aspects, such as the population by age and gender of France's communes.

# Introduction

Ce fichier est une base de données de l’INSEE paru le 27/02/2024 qui décrit les données sur plusieurs indicateurs. Cette base de données est encodée en UTF-8 sous format csv.

Il est séparé en deux fichiers, un fichier “dossier-complet” contenant plusieurs informations dans lequel chaque ligne représente une commune et un fichier “meta\_dossier\_complet” décrivant toutes les variables de notre fichier “dossier-complet” ainsi que toutes leurs modalités et leurs types.

Le fichier “dossier-complet” possède 35001 lignes et 1883 colonnes, où chaque colonnes représentent une variable.

Nous allons dans un premier temps vous présenter ces variables par thèmes puis vous expliquer le nettoyage éventuel à effectuer.

# Présentation des thèmes

## Évolution et structure de la population

Ce thème décrit l'évolution de la population en 2009, 2014 et 2020, dans lequel on peut les regrouper en plusieurs variables. Tout d'abord, ils ont réparti le nombre total de la population qui est écrit dans le fichier Excel sous cette forme : (P20\_POP), indiquant la population totale en 2020, où "P20" représente l'année et "POP" la population.

Ensuite, le nombre total de personnes réparties par tranche d'âge qui est sous cette forme : (P20\_POP1529), reprenant la forme de la population totale mais spécifiant une tranche d'âge de 15 à 29 ans.

Après, le nombre total de personnes est regroupé par genre : (P20\_POPH), indiquant le nombre total de la population par genre.

Par la suite, nous avons la combinaison de deux variables, celle de la population totale par tranche d'âge et celle de la population totale par genre : (P20\_F4559), indiquant le nombre total de femmes âgées de 45 à 59 ans.

Ensuite, nous avons la répartition du nombre total de personnes ayant habité dans le même logement, commune, département, région et outre-mer par tranches d'âge : (P20\_POP01P\_IRAN1), qui indique le nombre de personnes d'au moins 1 an ayant habité le même logement depuis 1 an en 2020, et (P20\_POP01P\_IRAN1), qui indique le nombre de personnes d'au moins 1 an ayant habité la même commune depuis 1 an en 2020, où "IRAN" indique la zone habitée par la personne. Enfin, ils ont réparti le nombre de naissances et de décès : (NAIS1420), indiquant le nombre de naissances entre janvier 2014 et janvier 2020, et (DECE1420), indiquant le nombre de décès entre janvier 2014 et janvier 2020.

## Caractéristique de l’emploi au sens du recensement

Le thème caractéristique de l’emploi au sens du recensement contient 345 variables réparties sur les années 2009, 2014, 2019 et 2020. Tout d’abord, il y a la variable "les actifs en 2020 de 15 à 64 ans (P20\_ACT1564) qui représente les personnes actives de 15 à 64 ans en 2020, dans laquelle "P20" indique l’année et (ACT1564) indique les actifs selon leur âge, ici de 15 à 64 ans.

Nous avons ensuite les actifs en 2020 par genre (P20\_HACT1564 et P20\_FACT1564). Ces variables représentent le nombre de personnes actives par genre. Nous avons également cette variable répartie selon le niveau de diplôme. Par exemple, les actifs qui ne possèdent aucun diplôme (P20\_ACT\_DIPLMIN). La variable actifs occupés de 15 à 64 ans (P20\_ACTOCC1564) est regroupée de la même façon que les actifs.

Ensuite, nous avons la variable Chômeurs de 15 à 64 ans en 2020 (P20\_CHOM1564). Nous avons une répartition de cette variable selon le niveau de diplôme, de BAC+2 à BAC+5. Par exemple, pour le BAC+5, nous avons (P20\_CHOM\_SUP5).

On a aussi une variable pour représenter les inactifs (P20\_INACT1564) qui représente les personnes inactives de 15 à 64 ans en 2020.

Il y a aussi une variable pour représenter les emplois salariés sur le lieu de travail (P20\_EMPLT) répartis selon leur contrat. Par exemple, (P20\_EMPLT\_SALTP) représente les emplois salariés sur le lieu de travail travaillant à temps partiel au lieu de travail en 2019.

Nous avons ces variables répétées pour les autres années (2009, 2014 et 2019).

## Couples - Familles - Ménages

Le thème Couples Familles Ménages contient 238 variables qui décrivent la situation familiale des individus, de 2009 à 2020 ayant 15 ans ou plus.

On a d’abord la “ situation matrimoniale des individus ayant 15 ou plus” avec par exemple le nombre de personnes de 15 ans ou plus mariées en 2020 (P20\_POP15P\_MARIEE), nombre de personnes de 15 ans ou plus pacsées en 2020 (P20\_POP15P\_PACSEE).

On a aussi cette variable pour les veufs, divorcée, célibataire et les personnes qui sont en concubinage ou en union libre.

Ensuite, il y a la variable “nombre de personnes des ménages”, on a par exemple la variable nombre total de personnes des ménages de 20 à 24 ans en 2020 (P20\_POPMEN2024).

On a aussi le nombre total de personnes par tranches d’âge vivant seules (P20\_POP1519\_PSEUL) représentant le nombre total de personnes vivant seules ayant un âge compris entre 15 et 19 ans en 2020.

Ensuite, nous avons “le nombre total de personnes par tranches d’âge vivant en couple” avec notamment le nombre total de personnes vivant en couple ayant un âge compris entre 15 et 19 ans en 2020 (P20\_POP1519\_COUPLE).

## Évolution et structure de l’entreprise

Ce thème possède 159 variables qui décrivent l’évolution des entreprises de 2013 à 2022.

Comme avec les précédents thèmes, on peut les regrouper en plusieurs catégories.

Tout d’abord, nous avons le nombre total d’entreprise, établissement et d’unités légales créées au cours d’une année, on utilisera comme exemple la variable (ENCTOT22) qui regroupe le nombre total d’entreprise créée en 2022.

Ensuite par le nombre total d’entreprises, d’établissements et d'unités légales créées par secteurs, dans lequel on utilisera par exemple (ENCBE22) qui regroupe le nombre total d’entreprises créer dans le secteur de l’industrie en 2022. Pour les établissements ça sera (ETCBE22) et unité légal (ENNBE22)

Le nombre total d'entreprises créées de manière individuelle au cours d'une année. On utilisera comme exemple (ENCITOT22), qui regroupe le nombre total d'entreprises créées par une seule personne en 2022.

Après, nous avons les mêmes variables mais cette fois-ci par secteurs, (ENCIBE22) qui regroupe le nombre total d’entreprises créées de manière individuel dans le secteur de l’industrie en 2022 (ETCBE22) pour les établissements dans le secteur de l’industrie et (ENNBE22) pour le nombre total d’unités légales.

Et en dernier, ils ont regroupé le nombre total d'établissements qui existe à la date du 31 décembre XXXX (ETNTOT22) qui regroupe le nombre d’établissements à la date du 31 décembre 2021.

Pareil pour le nombre d'établissements par secteur qui existe à la date du 31 décembre, (ETNBE22) qui regroupe le nombre total d’établissements dans le secteur de l’industrie à la date du 31 décembre 2021.

## Diplôme et Formation

Le thème Diplôme et Formation est basé sur les diplômes et formations des individus présents dans notre jeu de données répartis sur les années 2020 et 2009.

Tout d’abord, il y a la population qui est réparti sur 2 à 30 ans et plus, on a par exemple la caractérisation de la population entre 2 à 5 ans (P20\_POP0205) ou bien une tranche d'âge de 6 à 10 ans (P20\_POP0610), jusqu’à 30 ans et plus. Les tranches d'âge sont réparties entre 3 à 4 ans jusqu'à la dernière valeur où elle se base sur une tranche d'âge de 30 ou plus en 2020.

Puis, nous avons la population qui est scolarisée avec les mêmes tranches d'âge en 2020 que la variable précédente, on a par exemple les personnes qui sont scolarisées en 2020 qui ont entre 15 à 17 ans (P20\_SCOL1517).

Ensuite, nous avons une répartition pour cette variable basée sur les genres. On a par exemple le nombre d’hommes scolarisés de 25 à 29 ans en 2020 (P20\_HSCOL2529) pareil avec les femmes en 2020. On a aussi cette variable répartie avec les mêmes tranches d’âges que pour les autres variables comme par exemple (P20\_HSCOL1114) pour les hommes et (P20\_FSCOL1114) pour les femmes de 11 à 14 ans.

Ensuite, nous avons la population de 15 ans ou plus non scolarisée détenteur de diplôme ou non comme par exemple la population détenteur du brevet des collèges (P14\_NSCOL15P\_DIPLMIN), les détenteurs du bac (P14\_NSCOL15P\_BAC), ou bien encore des étudiants en enseignement supérieur (P14\_NSCOL15P\_SUP).

De plus, nous avons cette même répartition mais selon les genres.

Ensuite, il y a la population et la répartition des genres des personnes scolarisées en 2009, par exemple les hommes de 15 à 17 ans qui sont scolarisés (P09\_H1517).

Nous avons les mêmes variables pour l’année 2009 réparties de la même façon.

## Caractéristiques des établissements

Ce thème regroupe 95 variables qui décrivent les établissements fin 2021 dans lequel on peut les regrouper en plusieurs catégories.

Tout d’abord les établissements actifs en 2021 avec les établissements actifs de l'agriculture, sylviculture et pêche fin 2021 (ETAZ21**)** représentant le nombre d’établissements actifs qui travaille dans l’agriculture, sylviculture et pêche dans la fin de 2021et les établissements actifs dans divers secteur (industrie, commerce, construction).

Ensuite, il y a aussi les établissements actifs en 2021 par critère comme par exemple les établissements actifs sans salarié fin 2021 (ETTEF021) qui représente le nombre d’établissements actifs qui n’ont pas de salarié à la fin de l’année 2021.

Le nombre de postes des établissements est trié de la même façon que les établissements actifs.

Avec le nombre de postes des établissements actifs fin 2021 (ETPTOT21) représentant le nombre de postes disponibles dans les établissements actifs à la fin de l’année 2021.

Ensuite, ils ont regroupés par secteur. Il y a les établissements avec le nombre de postes des établissements actifs de l'agriculture, sylviculture et pêche fin 2021 (ETPAZ21) .

Regroupé par critère, il y a nombre de postes des établissements actifs de l'industrie de 50 à 99 salariés fin 2021 (ETPBE5021).

Il y a aussi la variable du nombre de particuliers employeurs d'assistants maternels fin 2021 (ETASSMAT21) qui représente le nombre de personnes qui sont employés dans l’assistance maternels dans la fin d’année 2021.

Et le nombre de particuliers employeurs autres activités en 2021 (ETAUTRES21) représentant le nombre de personnes employés dans toutes les autres activités dans la fin d’année 2021.

## Évolution des revenus et pauvreté des ménages

Ce thème contient 27 variables qui décrivent l’évolution des revenus et de la pauvreté des ménages.

D’abord, nous avons la variable nombre de ménages fiscaux avec le nombre de ménage fiscaux en 2021 (NBMENFISC2021) représentant le nombre total de ménages fiscaux en 2021.

Il y a aussi une variable similaire mais représentant le nombre total de personnes dans les ménages fiscaux (NBPERSMENFISC21).

Puis, nous avons la variable médiane du niveau de vie (MED21) qui représente le niveau de vie médian des ménages.

On a aussi une variable représentant la part des ménages fiscaux imposés (PIMP21).

Ensuite, nous avons la variable “taux de pauvreté” avec par exemple la variable (TP60AGE121) représentant le taux de pauvreté des moins de 30 ans en 2021.

Pour finir nous avons la variable décile du niveau de vie avec par exemple le 1er décile du niveau de vie en 2021 (D121).

## Naissances et décès domiciliés

Dans ce thème il n’y a que 17 variables qui décrivent les naissances et décès domiciliés de 2014 à 2022.

En premier, nous avons le nombre de naissances domiciliées au domicile de la mère, (NAISD16) qui regroupe le nombre de naissances domiciliées au domicile de la mère en 2016

Ensuite, nous avons le nombre de décès domiciliés au domicile du défunt de 2014 à 2022, (DECESD21) qui regroupe le nombre de décès domiciliés au domicile du défunt en 2021.

## Population active, emploi et chômage au sens du recensement

Dans ce thème, il y a 167 variables décrivant la population dans les années 2009, 2014, 2020 et 2021.

Nous avons la variable le nombre de personnes actives occupées de 15 ans ou plus en 2020 (P20\_ACTOCC15P) qui représente le nombre de personnes de 15 ans ou plus actives ayant un emploi en 2020.

Il y aussi le nombre actifs occupés de 15 ans ou plus en 2020 réparties par genre (P20\_HACTOCC15P) et par contrat d’entreprise par exemple (P20\_FACTOCC15P) qui représente les actifs occupés travaillant à temps partiel.

Nous avons aussi le nombre de personnes salariées de 15 ans ou plus en 2020 regroupées de la même façon que les personnes actives occupées, par exemple par genre (P20\_HSAL15P) et par contrat (P20\_HSAL15P\_TP). On a aussi une répartition de cette variable selon le genre et la tranche d’âge par exemple (P20\_FSAL1564\_TP) représentant les femmes salariées de 15 à 64 ans, il y a une répartition par tranches d’âge de 15 à 24 ans, 25 à 54 ans et 55 à 64 ans.

Il y a aussi la variable nombre de personnes non-salariées de 15 ans ou plus en 2020 représentant le nombre de personnes qui ne sont pas salariés et qui ont 15 ans ou plus en 2020 (P20\_NSAL15P) regroupés de la même façon que les personnes salariées, par contrat (P20\_NSAL15P\_TP) et genre (P20\_HSAL15P).

Puis, nous avons le nombre d'actifs occupés de 15 ans ou plus selon leur lieu de travail. Il y a par exemple le nombre d'actifs occupés de 15 ans ou plus qui travaillent dans la commune de résidence en 2020 (P20\_ACTOCC15P\_ILT1).

On a eu les mêmes variables pour les autres années.

On aussi le nombre d’actifs occupés de 15 ans ou plus, par différence de lieu en 2009 par exemple nombre d'actifs occupés de 15 ans ou plus qui travaillent dans dans une autre commune située dans une autre région en France métropolitaine en 2009 (P09\_ACTOCC15P\_ILT4).

## Salaire et revenus d’activité

Dans ce thème, il y a 25 variables qui décrivent les salaires et revenus de la population pour l'année 2021. Premièrement, il y a le salaire net moyen généré par la population dans son ensemble (SNHM21). Par la suite, les variables concernent les cadres, les professions intellectuelles ou les chefs d'entreprise en fonction de leurs revenus (SNHMC21).

Ensuite, les salaires concernent l’activité professionnelle de la population, comme les ouvriers (SNHMO21), et les employés en général (SNHME21).

De plus, nous avons une répartition des salaires selon le genre par rapport à l’activité professionnelle, avec (SNHMFO21) pour les femmes ouvrières et (SNHMHO21) pour les hommes ouvriers en 2021.

Enfin, nous avons une répartition des salaires selon des tranches d'âge de la population, avec (SNHM1821) pour les salaires nets des personnes âgées de 18 à 25 ans en 2021. Cela concerne également tous les âges ; il existe également une répartition similaire basée sur les âges et les genres.

## Tourisme

Dans ce thème, il y a 52 variables qui décrivent les emplacements touristiques. Premièrement, il y a les hôtels en 2024 qui sont également répartis selon le niveau de qualité des hôtels, par exemple les hôtels 2 étoiles en 2024 (HT224). Ensuite, il y a les chambres des hôtels selon la qualité des hôtels, par exemple une chambre d'hôtel 5 étoiles (HTCH524).

Par la suite, il y a les campings répartis selon leur qualité, soit par exemple un camping 4 étoiles (CPG424). De plus, il y a la location d'un emplacement à l'année dans le camping, toujours basée sur le critère de qualité, comme par exemple un emplacement de camping loué dans un camping 3 étoiles (CPGEL324). Toujours concernant les campings, il y a les emplacements offerts, comme un emplacement offert dans un camping 1 étoile (CPGEO124).

Enfin, il y a les villages vacances, notamment dans l'hébergement (VVUH24), les lits (VVLIT24) et la résidence tourisme (RT24), qui sont également présents pour la jeunesse avec les mêmes types d'établissements, comme l'auberge de jeunesse - centre sportif en 2024 (AJCS24).

## Logement

Ce thème contient 258 variables décrivant les différents types de logements.

D'abord, il y a la variable du type d'habitation, avec par exemple le nombre de maisons en 2020 (P20\_MAISON).

Ensuite, nous avons le nombre de résidences construites entre plusieurs années jusqu'en 2020, avec par exemple (P20\_MAP20\_RPMAISON\_ACH90) représentant le nombre de résidences principales de type maison construites de 1971 à 1990 en 2020.

Nous avons le nombre de ménages ayant emménagé depuis un certain temps, par exemple (P20\_PMEN\_ANEM0204) représentant la population des ménages ayant emménagé entre 2 et 4 ans en 2020.

Ensuite, nous avons le nombre de résidences principales occupées par une fonction, avec par exemple (P20\_RP\_PROP) représentant le nombre de résidences principales occupées par des propriétaires en 2020.

On a aussi le nombre total d’ancienneté d’emménagement dans les résidences principales (P20\_ANEM\_RP\_PROP) représentant l’ancienneté totale d'emménagement dans les résidences principales occupées par des propriétaires en années en 2020.

Ensuite, nous avons le nombre de résidences possédant un bien, avec par exemple (P20\_RP\_BDWC) représentant le nombre de résidences principales avec baignoire ou douche et WC à l'intérieur du logement en 2020, et (P20\_RP\_VOIT1) représentant le nombre de ménages disposant d'une voiture en 2020.

# Nettoyage éventuel

Après une ouverture du fichier avec la librairie “Pandas” et une rapide exploration dans lequel, nous pouvons constater :

* Des lignes avec de nombreuses colonnes vides.
* Des colonnes avec de nombreuses lignes vides.
* Certaines colonnes ne sont pas au bon format ( ex : au lieu d'être des entiers, c’est des float).
* Dans certaines colonnes, le séparateur des décimales est la virgule au lieu du point
* Il y a des cellules qui contiennent des chaînes de caractères.

Ainsi pour résoudre des problèmes, nous allons :

* Supprimer toutes les colonnes avec trop de lignes vides
* Supprimer toutes les lignes avec trop de colonnes vides
* Remplacer toutes les virgules par des points (chiffres)
* Remplacer toutes les chaînes de caractères par des 0
* Remplir toutes les cellules vides restantes par des 0
* Convertir (cast) chaque colonne dans le bon format. pour ce cas, nous attendrons après avoir sélectionné les variable que l’on va utiliser avant de les caster en entier, car il y’a certain variable qui sont des taux (pourcentage)
* enfin, créer une nouvelle colonne ‘DEPARTEMENT’ qui nous permettra de regrouper les donnée

# Conclusion

Pour conclure cette étape, qui consistait à faire une exploration des données présentes sur le fichier de l’INSEE notamment sur les variables présentes.

Nous pouvons noter que le fichier télécharger sur le site de l’INSEE était au format zip et contient 2 fichiers au format csv, meta\_data et data. data contient les données proprement parler, quand a meta\_data, il décrivait tous les variable présente dans data, on avait :

* code\_var : le code attribué à la variable
* lib\_var : le libellé de la variable (version courte)
* lib\_var\_long : le libellé de la variable (version longue)
* cod\_mod : le code attribué à chaque modalité de chaque variable
* lib\_mod : le libellé de chaque modalité
* type\_var : le type de chaque variable
* long var : la longueur de chaque variable
* thème : le thème auquel appartient la variable

Toutes les variables sont regroupées par thème et sont au nombre de 12.

Ces thèmes abordent différents sujets tels que la démographie, le chômage, le tourisme, etc…

Dans le fichier data, chaque colonne représente une variable et chaque ligne représente une commune. nous avons 35001 lignes et 1883 colonnes.

par la suite, nous avons procéder au nettoyage des donnée à l’aide de scripts python permettant de supprimer les colonnes inintéressantes, c’est-à-dire celles qui contenaient trop de valeurs vides, mais aussi de corriger les valeurs incorrectes (supprimer les chaînes de caractères des cellule car toutes nos données sont des valeurs numériques), et remplissage des cellules vides restantes par des 0.

# Bibliographie

Insee : [Base du dossier complet | Insee](https://www.insee.fr/fr/statistiques/5359146#consulter)