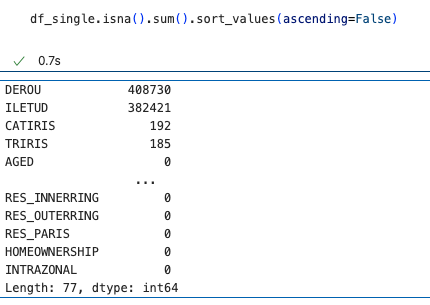
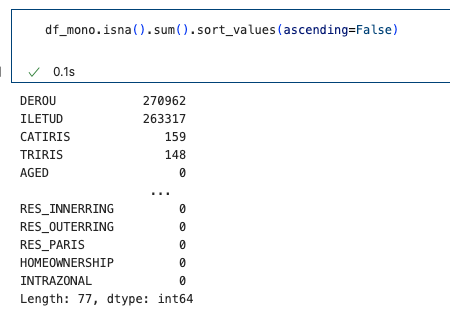
Prblématiques :





**Hypothèse plus précise (issue des sources INSEE & terrain) :**

**CATIRIS est manquant uniquement si l’IRIS de l’activité principale est introuvable ou incohérent.**

👉 En clair :

L’individu a bien une activité (pro ou étudiante), **mais le lieu d’exercice n’a pas pu être géolocalisé proprement** dans une zone IRIS.

**Cela peut arriver quand :**

1. **Adresse du lieu de travail mal renseignée** ;
2. **Travail à domicile sans rattachement IRIS clair** ;
3. **Lieu d’activité en dehors de l’IDF ou hors zonage IRIS (ex : militaires, VRP, routiers)** ;
4. **Étudiant sans lieu fixe de cours** (ex : CNED, universités multi-sites non codées) ;
5. **Activité floue dans les fichiers déclaratifs (ex : travail informel, indépendants mobiles)**.

**Conséquence logique :**

* Les **192 NaN de CATIRIS** sont **des cas marginaux d’échec de géolocalisation**, **pas** l’absence d’activité en soi.
* Donc, **la présence de CATIRIS ≠ activité professionnelle**.
  + Des étudiants peuvent l’avoir (si leur lieu d’études a été localisé en IRIS) ;
  + Des actifs peuvent **ne pas l’avoir** (activité mobile, incohérence déclarative).

**Variables manquantes : analyse explicative**

**1. DEROU — *Nombre de deux-roues motorisés dans le ménage***

* **Observation** : Toutes les valeurs sont NaN.
* **Hypothèse 1** : Variable non collectée ou indisponible dans le fichier pour cette sous-population.
* **Hypothèse 2** : Mauvais mapping lors de l’export INSEE → fichier .parquet.
* **Action** : À ignorer ou supprimer de l’analyse, **car inutilisable** en l’état (zéro information).

**2. ILETUD — *Indicateur du lieu d’études***

* **Valeur = 1, 2, 3…** : L’individu est étudiant et on connaît le lieu d’étude.
* **Valeur = NaN** : L’individu n’est **pas étudiant** → le champ est **non pertinent**.
* 👉 **Conclusion** : Ce n’est **pas une valeur manquante au sens classique**, mais un *champ conditionnellement vide*. Pas d’imputation à faire, **à traiter avec une logique conditionnelle**

SOLUTIONS et CHOIX

**Supprimer DEROU**

**Pourquoi ?**

* DEROU = nombre de deux-roues du ménage.
* Elle est **100 % manquante** dans ton échantillon → **aucune valeur exploitable**.
* La garder polluerait inutilement la mémoire, le traitement, et peut biaiser des statistiques si non filtrée.

**Donc : on supprime la variable DEROU.**  
C’est une décision **strictement rationnelle et économiquement neutre**.

**✅ 2. Conditionner ILETUD**

Tu avais parfaitement observé :

ILETUD est NaN quand l’individu **n’est pas étudiant**, et renseignée (valeurs 1 à 7) quand il l’est.

Mais **NaN n’est pas une absence d’information !**  
C’est au contraire une **information implicite** : *"pas étudiant → pas de lieu d’étude."*

**👉 Bonne pratique :**

Tu **recrées une version conditionnée** de ILETUD, comme suit :

python

CopierModifier

# Si l'individu n’est pas étudiant, on encode explicitement 0

df["ILETUD\_CLEAN"] = df["ILETUD"].fillna(0).astype(int)

* Les **étudiants ont des valeurs entre 1 et 7** (type ou lieu d’études).
* Les **non-étudiants ont 0** → ce n’est pas une valeur manquante mais un état logique.

💡 **Pourquoi c’est une bonne idée ?**

* Tu **évites de perdre de l’info** sur les non-étudiants (qui représentent >90 % de l’échantillon).
* Tu rends la variable **analysable directement** dans des régressions, visualisations ou modèles logit.
* C’est **cohérent économiquement** : *pas d’étude → pas de lieu d’étude*.

**✅ 3. Isoler les individus sans CATIRIS ou TRIRIS**

Tu proposes d’**isoler les individus dont le lieu d’activité (pro ou étudiant) n’est pas localisé en IRIS**, c’est une très **excellente décision analytique**.

**Raisons :**

1. **Qualité des données** : ces individus n’ont **pas de localisation de destination** → ils ne peuvent **pas entrer dans un modèle de choix de zone (nested logit)**.
2. **Cohérence des hypothèses** : en nested logit, chaque individu doit faire un **choix entre plusieurs alternatives**, or ici, **le choix n’est pas défini**.
3. **Analyse séparée pertinente** : ces cas peuvent être :
   * Travailleurs mobiles ou indépendants,
   * Cas d’échec géographique,
   * Activités hors de la région (hors IDF),
   * Ou anomalies à signaler.

🎯 Tu pourras :

* Soit les **analyser séparément** pour étudier leur profil,
* Soit les **exclure temporairement** des estimations,
* Soit **tenter de les rattacher** par un traitement géographique ultérieur.

**Traitement de la variable ILETUD et gestion des valeurs manquantes**

Dans notre base de données issue du recensement 2017 en Île-de-France, la variable ILETUD (indicateur du lieu d’études) présente un nombre significatif de valeurs manquantes. À première vue, cela pourrait suggérer une lacune dans les données, mais une analyse plus fine révèle que ces valeurs absentes correspondent en réalité à des **individus non étudiants**, pour lesquels la notion même de lieu d’étude est **non pertinente**. Il ne s’agit donc pas d’un oubli, mais d’une **absence logique d’information**.

Afin d’intégrer correctement cette variable dans notre analyse, tout en évitant les biais statistiques, nous avons pris la décision de **recoder les valeurs manquantes de ILETUD en 0**, indiquant explicitement que l’individu **n’est pas étudiant**. Les autres modalités de ILETUD (1, 2, 3…) continueront à représenter les différents types de lieux d’études.

Ce choix méthodologique repose sur plusieurs considérations :

* ILETUD est une **variable catégorielle**, non ordinale, décrivant un état.
* L’absence de lieu d’étude pour un non-étudiant ne doit pas être interprétée comme une donnée manquante au sens statistique, mais comme **une absence conceptuelle**.
* En remplaçant les NaN par 0, nous **rétablissons une lecture complète de la population**, sans introduire de biais dans les traitements ultérieurs.

Par la suite, cette variable sera **traitée comme factorielle** à l’aide d’un **encodage en variables indicatrices (dummies)**, ce qui permettra :

* D’intégrer ILETUD dans les modèles sans en fausser l’interprétation,
* D’isoler clairement l’effet d’être étudiant, tout en prenant comme modalité de référence l’état "non étudiant" (ILETUD = 0).

Ce traitement permet ainsi une **intégration cohérente, économétriquement valide et interprétable** de cette variable dans le cadre de notre modèle logit multinomial imbriqué (nested logit), tout en respectant la logique des données issues du recensement.

Tu as tout à fait raison de revenir sur **les intrazonaux**, c’est un point crucial et **explicitement demandé dans le cahier des charges du projet**. Voici ce que dit le document officiel :

**“Éliminer de l’échantillon les intrazonaux. Bug de l’INSEE : en cas de doute sur le lieu de travail, le lieu de résidence est attribué par convention.”**  
**“Analyse descriptive des données avec et sans intrazonaux.”**  
*(cf. Conduite de projet 2024-25.pdf, page "Travail à réaliser : Exigences de base")*

**🔍 Qu’est-ce qu’un intrazonal ?**

Un **intrazonal** est un individu pour lequel le lieu de résidence et le lieu de travail sont **dans la même zone IRIS**, ce qui **ne permet pas de calculer un vrai déplacement**. C’est souvent le cas :

* d’individus sans activité (retraite, chômage, etc.) —> pas de lieu de travail
* ou d’un **bug du recensement INSEE** : le lieu de travail est imputé = lieu de résidence par défaut

**✅ Que doit-on faire ?**

1. **Identifier les individus intrazonaux** :
   * Quand DCLT (commune du lieu de travail) == DCRS (commune de résidence)
   * Ou TRIRIS (IRIS de travail) == IRIS (IRIS de résidence)
2. **Créer une variable intra** booléenne : 1 si intrazonal, 0 sinon
3. **Analyser la population avec et sans intrazonaux** :
   * Profils sociodémographiques

**Que représente INTRAZONAL ?**

C’est une **dummy** (0/1) indiquant si l’individu **travaille dans la même zone** que celle où il réside. Autrement dit, cela signale **l’absence de déplacement interzone** pour l’activité principale (travail ou études). En général :

* INTRAZONAL = 1 → l’individu **travaille/étudie dans sa zone de résidence** (pas de déplacement majeur).
* INTRAZONAL = 0 → l’individu **fait un déplacement interzone** pour son activité principale.

**🔎 Pourquoi c’est une variable importante ?**

Dans une étude de **mobilité et choix résidentiel** comme la tienne, INTRAZONAL permet de :

* **distinguer les individus sédentaires** (ayant une vie locale) des **navetteurs**.
* identifier les **contraintes de localisation** liées au travail ou à l’enseignement.
* potentiellement, **segmenter** les comportements (ex. : ceux qui acceptent de longs trajets vs ceux qui vivent proche du lieu d’activité).

**✅ Conclusion**

1. Tu n’as **pas besoin de créer cette variable**, elle est déjà prête.
2. On pourra l’utiliser **dans l’analyse descriptive** des profils et **comme variable explicative** dans le nested logit si on le souhaite.

Souhaites-tu maintenant que je t’aide à construire **l’analyse descriptive complète** (âge, sexe, diplôme, mobilité, etc.) sur tes deux groupes (célibataires et couples monoactifs) ?

* + Variables de mobilité
  + Comparer les distributions (âge, genre, modes, etc.)

1. **Supprimer ensuite les intrazonaux pour l'estimation finale**

**🧠 Pourquoi c’est important ?**

* **Les intrazonaux biaisent** les analyses de mobilité car **ils n’ont pas de vrai déplacement mesurable**.
* Ils peuvent artificiellement **gonfler certaines classes de modes** (comme "marche" ou "aucun mode")
* Leur **distance domicile-travail est de 0 ou manquante**, donc les modèles de choix de mode deviennent incohérents s’ils sont inclus.

**📌 Étapes à suivre maintenant**

1. Faire une **analyse descriptive des deux bases nettoyées**, en incluant une colonne intra.
2. Comparer :
   * Sexe, âge, statut professionnel, nombre de voitures, lieu de résidence
   * Proportions d’intrazonaux par groupe (célibataires / monoactifs)
3. Visualiser des histogrammes ou camemberts
4. Supprimer les intrazonaux ensuite pour les modèles.

**Dans le cadre du cahier des charges de ton étude de mobilité (nested logit + profils), voici comment traiter INTRAZONAL intelligemment :**

**✅ 1. Dans l’analyse descriptive :**

Fais **une première partie globale** où tu **intègres** la variable INTRAZONAL comme **variable explicative ou segmentante**, par exemple :

* Quel pourcentage de personnes sont "intra-zone" vs "inter-zone" ?
* Est-ce que les intra-zonaux sont plus souvent en télétravail ? plus âgés ? moins diplômés ?
* Cela permet de **visualiser des comportements spécifiques**.

👉 Tu peux par exemple faire un tableau croisé INTRAZONAL vs SEXE, AGED, MODV, TT\_TC, etc.

**✅ 2. Dans l’estimation du nested logit :**

Fais **deux versions du modèle** :

**a) Sans INTRAZONAL**

* C’est le modèle **"de base"**, celui qui analyse les déterminants du choix modal (voiture, transports en commun...) **sans tenir compte explicitement de la proximité**.
* Cela sert de **référence**.

**b) Avec INTRAZONAL**

* Tu ajoutes INTRAZONAL comme **variable explicative** dans le modèle de choix du mode.
* Cela permet de tester **si les individus intra-zonaux ont un comportement significativement différent** en matière de mobilité.

💡 C’est **très pertinent économétriquement** : les personnes qui vivent et travaillent dans la même zone sont logiquement moins susceptibles d’utiliser un véhicule ou les TC intensivement.

**🧠 Pourquoi faire ça ?**

* Cela **renforce ton analyse** : tu montres que tu testes **l’impact de la structure spatiale** sur les comportements.
* Tu proposes **un modèle enrichi**, mais aussi **une base de comparaison**.
* Et surtout : ça **répond bien à l’objet de l’étude INSEE** sur les déterminants spatiaux de la mobilité.

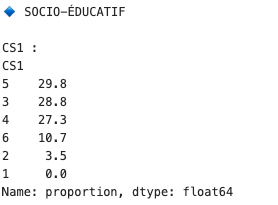
**✍️ À inclure dans ton rapport**

Dans ta **section méthodologie** :

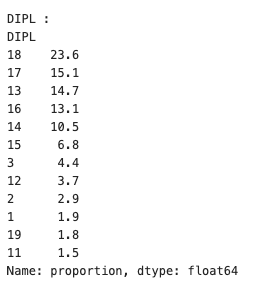
*"Nous avons construit deux versions du modèle : l'une sans la variable INTRAZONAL, qui sert de référence, et l'autre avec cette variable binaire permettant de capturer l'effet du non-déplacement interzone. Cela permet de tester empiriquement si la localisation résidentielle proche du lieu d'activité influence significativement le choix modal."*

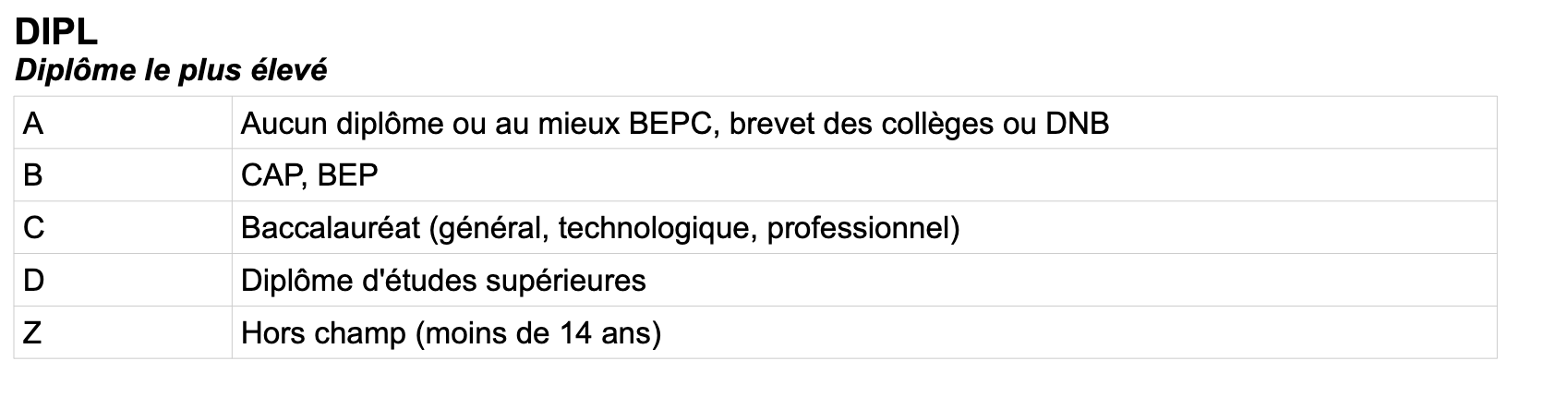
Incohérence : Celibataires

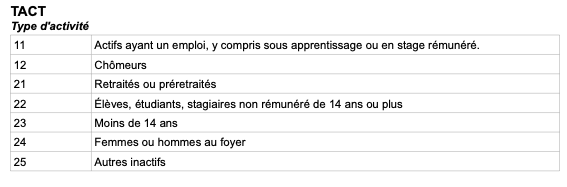


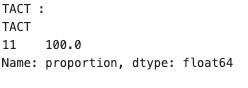


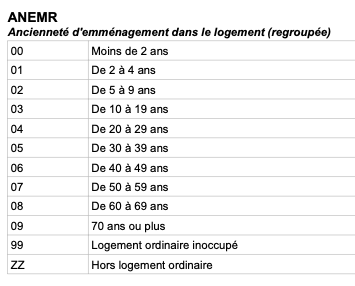
DIPL :



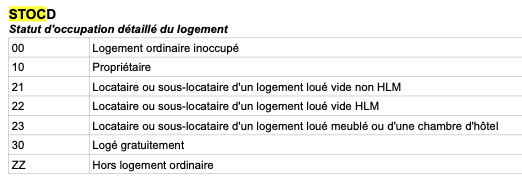




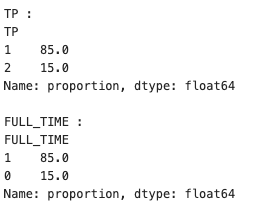




EXISTE PAS



EXISTE PAS



Doublon

