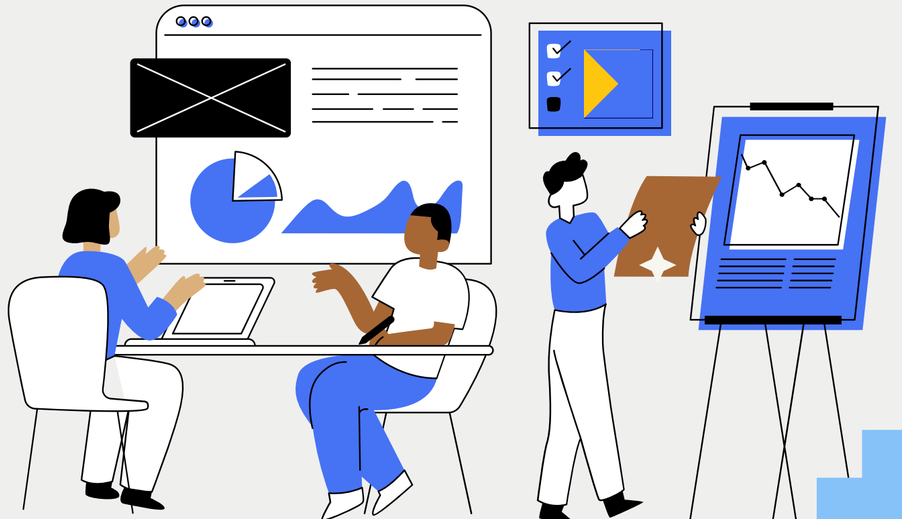
DU DATA ANALYTICS

8 juillet 2024

Projets Statistiques



Etudiants :
CAMARA Zakaria
CHAPELLON Nolwenn
SALMON Joris

SOMMAIRE

Régression logistique

1 Construction du jeu de donnée

CHOIX DES VARIABLES : À LA MAIN, SÉLÉCTION D'UNE TRENTAINE DE VARIABLE PERTINANTE AVEC LE SUJET

```
colonne = [
    'IDENT_MEN', 'IDENT_IND', 'NPERS', 'NACTIFS', 'TYPMEN5', 'REVMEN',
    'CJSITUA', 'CJACTOCCUP', 'SEXE', 'AGE', 'COUPLE', 'COUPLRP', 'CONJOINT',
    'ETAMATRI', 'ENFANT', 'IPROPLOC', 'PRACT', 'LIENPREF', 'SITUA',
    'RABS', 'NBENFM3', 'NBENF3A17', 'NBENF18P', 'TOTREVEN', 'ITOTREV', 'EREG',
    'FRANCE', 'TPP', 'HH', 'ACTIP', 'STATUTP', 'ACTIM', 'STATUTM',
    'ETUDIPL', 'STATUTEXT', 'STATUT', 'CLASSIF', 'TYPEMPLOI', 'FONCTION',
    'ETUDES', 'DIPLOME', 'NATIO', 'AGFINETU', 'NBSALENTC', 'IDENT_LOG',
    'pondcal', 'naf17', 'naf4', 'region', 'anciennete', 'naf38', 'nati', "Y"
]
```

FILTRER LES INDIVIDUS VIVANT EN COUPLE DANS LE MÊME LOGEMENT : ON CONSTRUIT LA VARIABLE BINAIRE

```
# Filtre des individus vivant en couple dans le même logement
df = df[df['COUPLE'] == "1"]

# Variable binaire
df['Y'] = df.apply(lambda row: 1 if (row['SEXE'] == '2') & (row['PRACT'] == '1') else 0, axis=1)
```

2 Analyse statistique

SEPARATION DES VARIABLES QUALITATIVES ET QUANTITATIVES

```
# Listes des variables quantitatives et qualitatives
quanti_cols = ['NPERS', 'NACTIFS', 'REVMEN', 'AGE', 'NBENFM3', 'NBENF3A17', 'TOTREVEN', 'HH', 'AGFINETU', 'pondcal', 'anciennete']
quali_cols = ['TYPMEN5', 'CJSITUA', 'CJACTOCCUP', 'COUPLRP', 'CONJOINT', 'ETAMATRI', 'ENFANT', 'RABS','IPROPLOC', 'LIENPREF', 'SITUA'
```

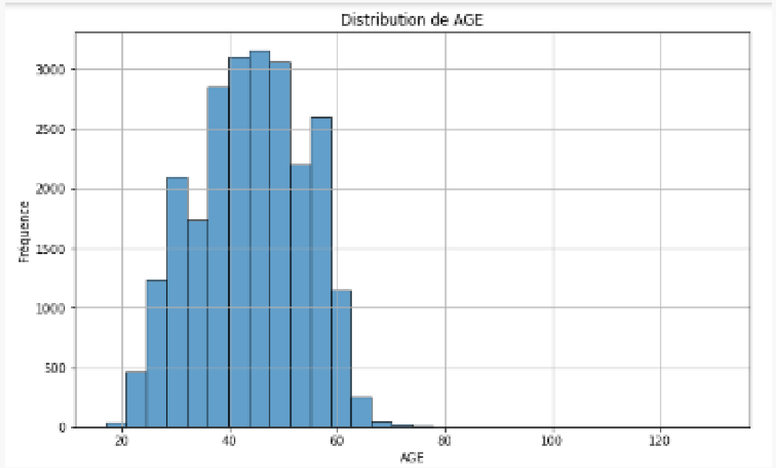
ANALYSE UNIVARIEE :

STATISTIQUES DESCRIPTIVES

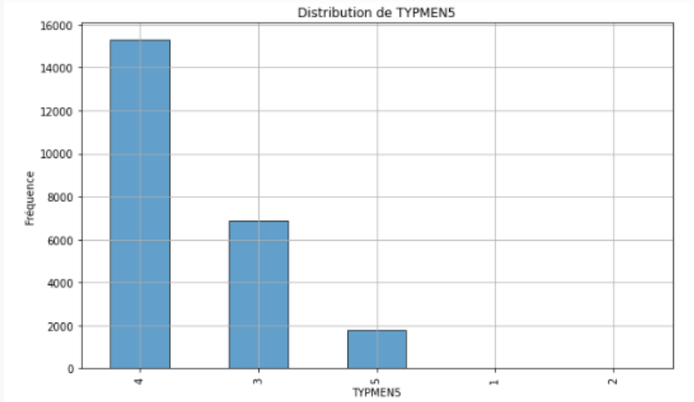
```
selected_df.describe()
```

	NPERS	NACTIFS	REVMEN	CJACTOCCUP	AGE
count	24024.000000	24024.000000	22989.000000	24020.000000	24024.000000
mean	3.337829	1.956252	3821.640045	0.792381	43.817974
std	1.196510	0.523638	2724.325930	0.405611	10.094510
min	0.000000	0.000000	100.000000	0.000000	17.000000
25%	2.000000	2.000000	2600.000000	1.000000	36.000000
50%	3.000000	2.000000	3500.000000	1.000000	44.000000
75%	4.000000	2.000000	4500.000000	1.000000	52.000000
max	18.000000	7.000000	95000.000000	1.000000	131.000000

HISTOGRAMME POUR LA VARIABLE QUANTITATIVE AGE (NBRE DE PERSONNES DU MÉNAGE)



BAR PLOT POUR LA VARIABLE QUALITATIVE TYPMEN5 (TYPE DE MÉNAGE)



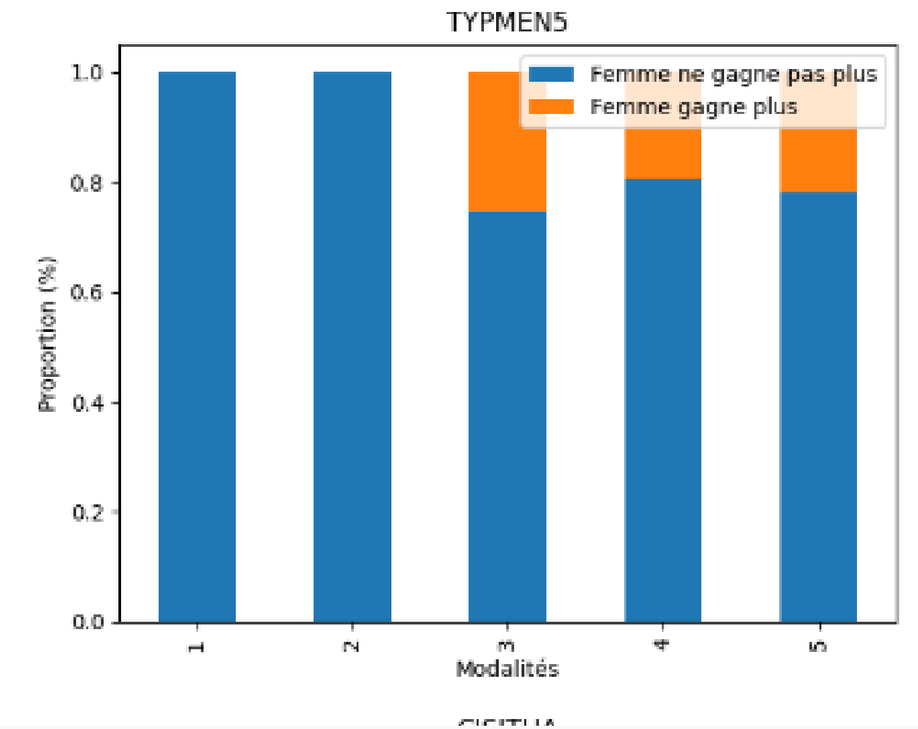
2 Analyse statistique

SEPARATION DES VARIABLES QUALITATIVES ET QUANTITATIVES

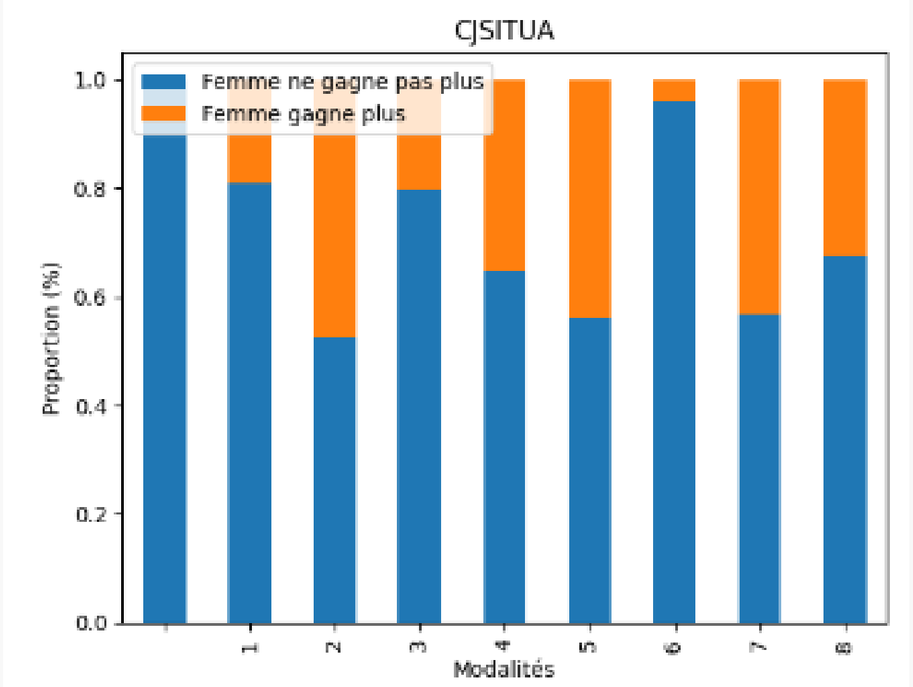
```
# Listes des variables quantitatives et qualitatives
quanti_cols = ['NPERS', 'NACTIFS', 'REVMEN', 'AGE', 'NBENFM3', 'NBENF3A17', 'TOTREVEN', 'HH', 'AGFINETU', 'pondcal', 'anciennete']
quali_cols = ['TYPMEN5', 'CJSITUA', 'CJACTOCCUP', 'COUPLRP', 'CONJOINT', 'ETAMATRI', 'ENFANT', 'RABS','IPROPLOC', 'LIENPREF', 'SITUA'
```

ANALYSE BIVARIEE

GRAPHIQUE BIVARIEE POUR LA VARIABLE QUALITATIVE TYPMEN5 (TYPE DE MÉNAGE)

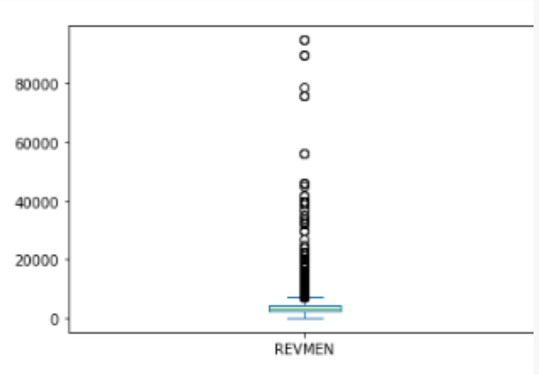


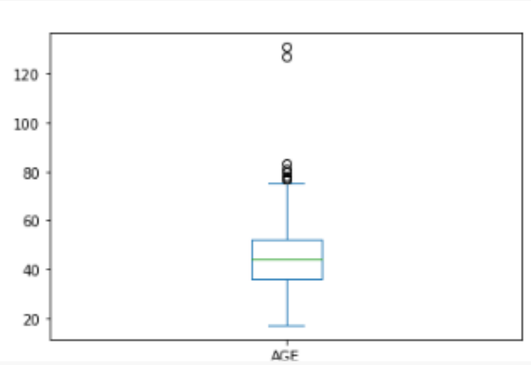
GRAPHIQUE BIVARIEE POUR LA VARIABLE QUALITATIVE CJSITUA (SITUATION PRINCIPALE DU CONJOINT DU L)



3 Gestion des NA

OUTLIERS : Visualisation et Traitement des variables





REVMEN et HH (Revenu mensuel/Indicateur de chef de ménage)

- Remplacer les NA en fonction de la médiane des groupes de SITUA

CJACTOCCUP (Conjoint actif occupé)

- Calculer le mode et remplacer les valeurs NA par le mode

AGFINETU (Âge de fin d'études)

- Remplacer les valeurs NA de AGFINETU par la médiane arrondie de AGFINETU en fonction de DIPLOME

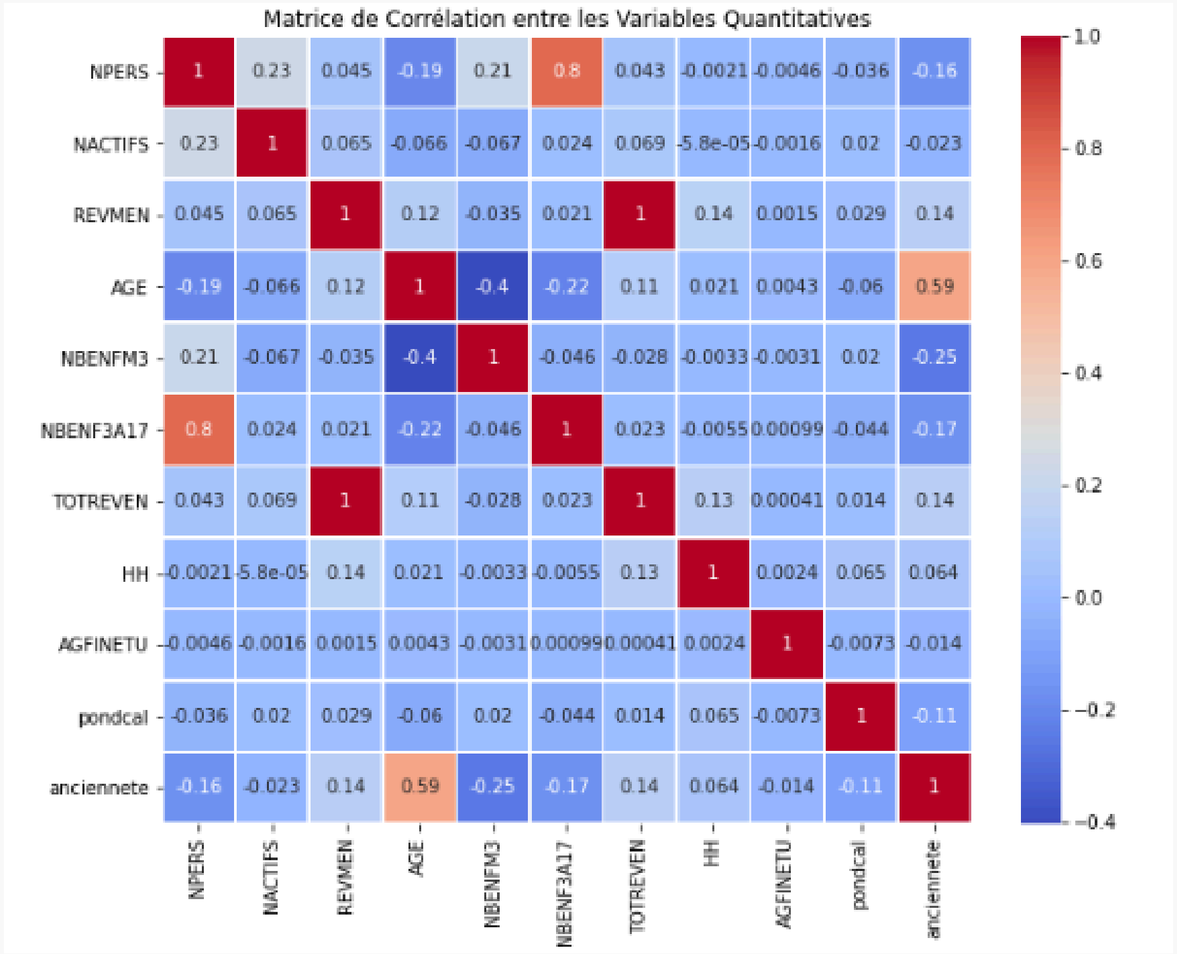
ANCIENNETE (Ancienneté dans l'emploi)

- Remplacer les valeurs NA de "anciennete" par la moyenne de "anciennete" en fonction de "AGE"

RABS (Raison de non travail)

- NA pour les personnes, on remplace alors les NA par 0

MATRICE DE CORRÉLATION

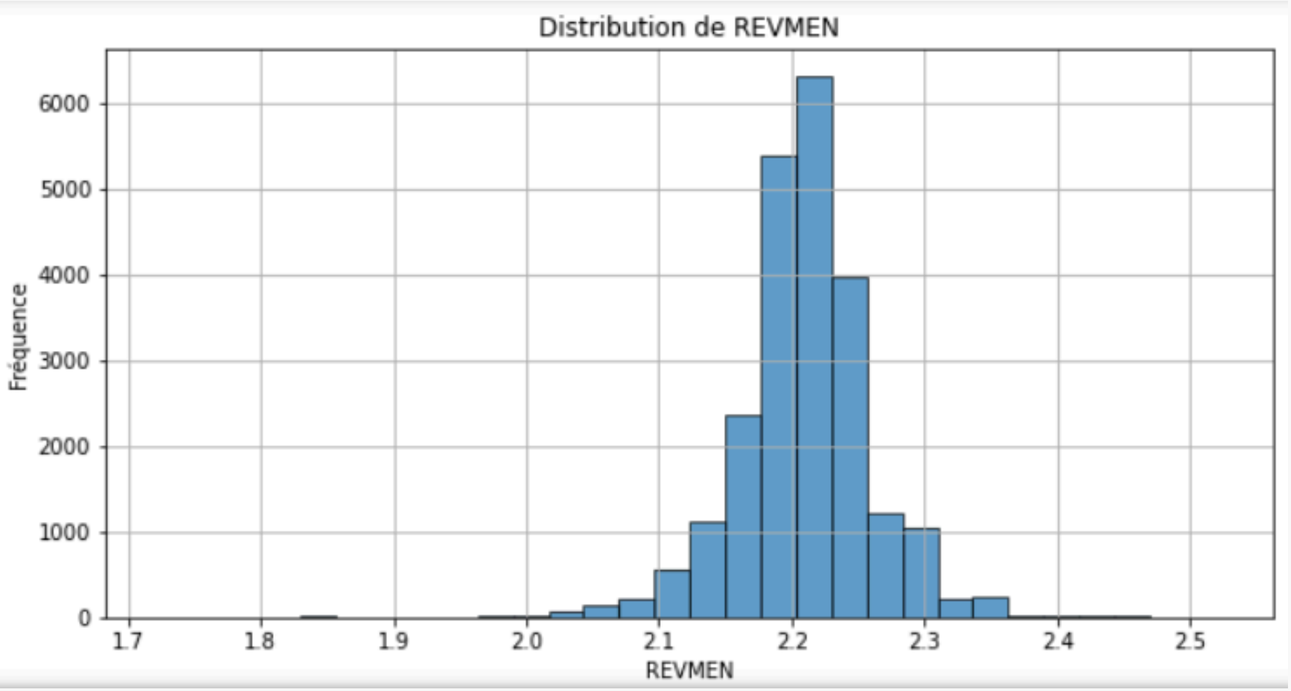


Analyse statistique

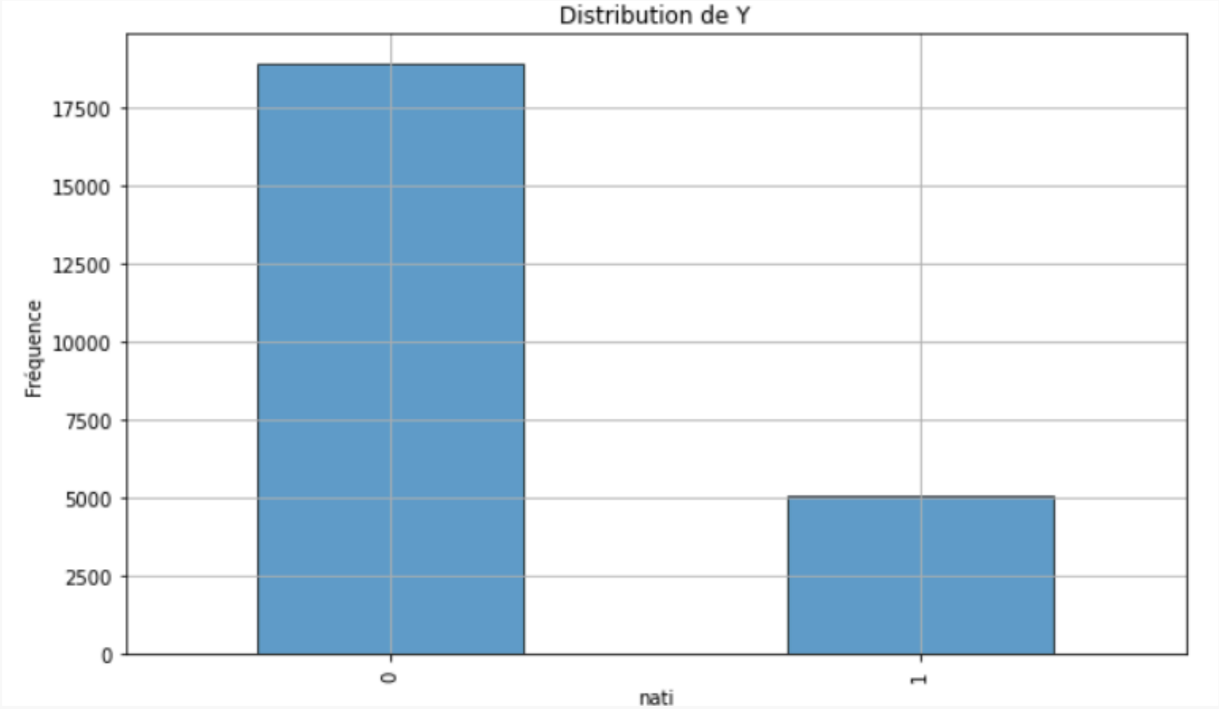
TRANSFORMATION

Normalisation de la variable quantitative “REVMEN” qui suit une loi de puissance avec le log

REVMEN APRÈS TRANSFORMATION EN LOG



DISTRIBUTION DES 0 ET 1 DE LA VARIABLE Y.



Déséquilibre des classes : On peut voir un déséquilibre entre dans la distribution

Régression logistique

Sélection des variables :

- Filtrer les variables à faible variance
- Enlever les coefficients inférieurs à un certain seuil

	Modèle 1	**Modèle 2**
Méthode	Forward	Backward
f1-Score	0,75	0,75
AIC	7484.37	7383.34
BIC	7738.55	7637.52

Variable déterminante

FONCTION_3 (Odds Ratio = 0.5636):

Les femmes ayant un métier dans la catégorie “Gardiennage, nettoyage, entretien ménager” ont environ 44% moins de chances d'être le principal apporteur de ressources.

CJSITUA_6 (Odds Ratio = 0.5739):

La situation de “Femme foyer” conduit à une réduction de 42% de la probabilité qu'une femme soit le principal apporteur de ressources.

region_11 (Odds Ratio = 1.4392):

Les femmes vivant dans la région “Île-de-France” ont environ 1.44 fois plus de chances d'être le principal apporteur de ressources.

naf38_CH (Odds Ratio = 1.3376):

Les femmes dans cette catégorie d'activité (Métallurgie & fab. de prdts métalliques sauf machines & équipmnts) ont environ 1.34 fois plus de chances d'être le principal apporteur de ressources.

CJSITUA_5 (Odds Ratio = 0.7527):

La situation de “Retraité(e) ou retiré(e) des affaires ou en préretraite” conduit à une réduction de 25% de la probabilité qu'une femme soit le principal apporteur de ressources.

Décrochage scolaire

Préparation des données

1) Sélection des colonnes

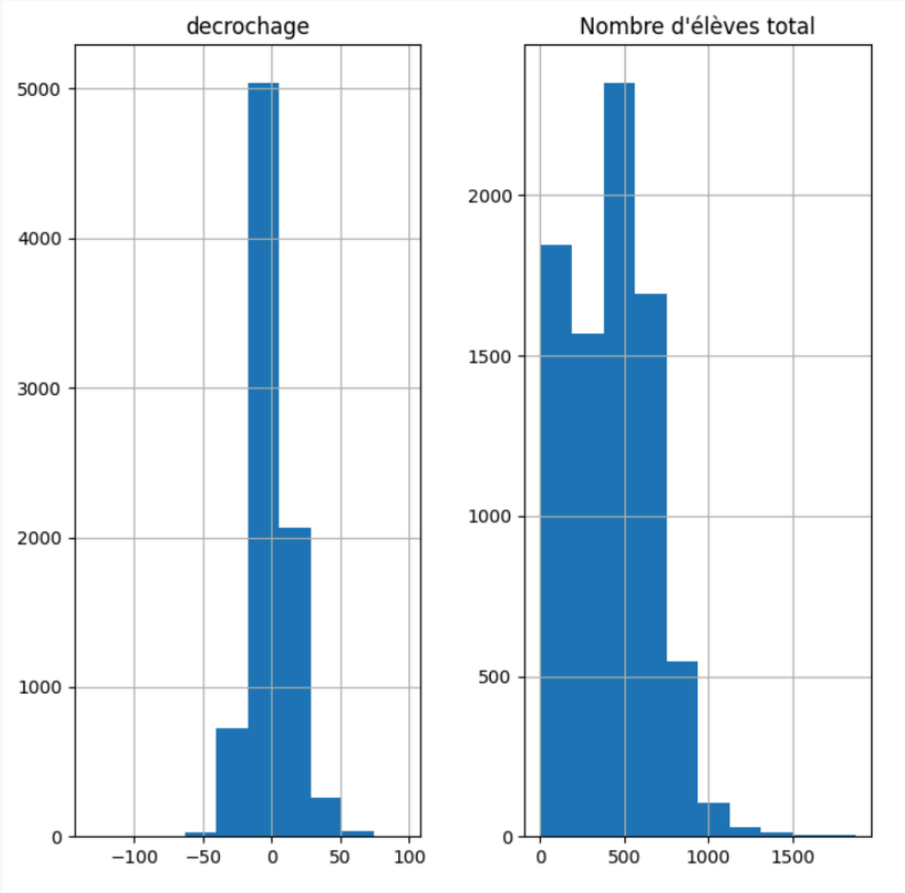
Essentielles pour répondre aux questions

2) Pré-calcul de certaines variables

Décrochages (nb élèves 3ème -6ème -> entre 2019 et 2022)

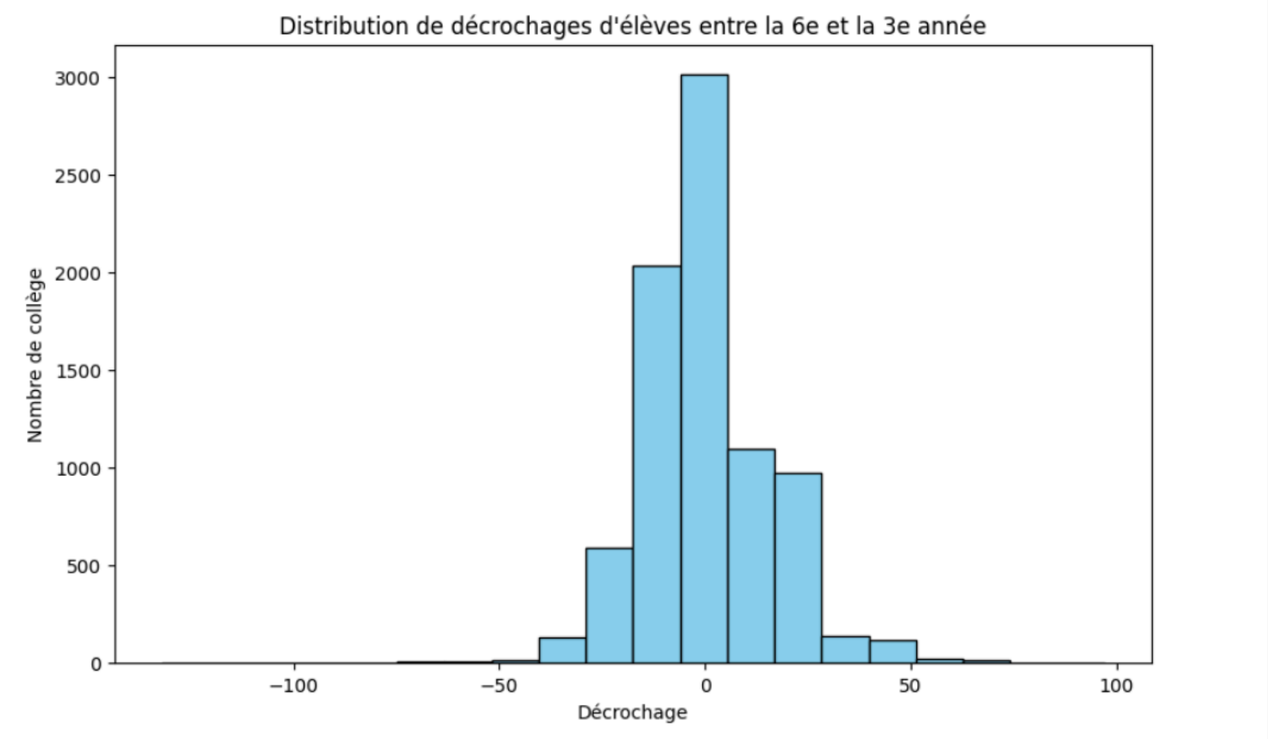
Décrochages (filles & garçons)

4) Visualisation de la distribution de nos données



Question

Question 1 : Certains collèges ont t'ils davantage de décrochages scolaire que d'autres ?



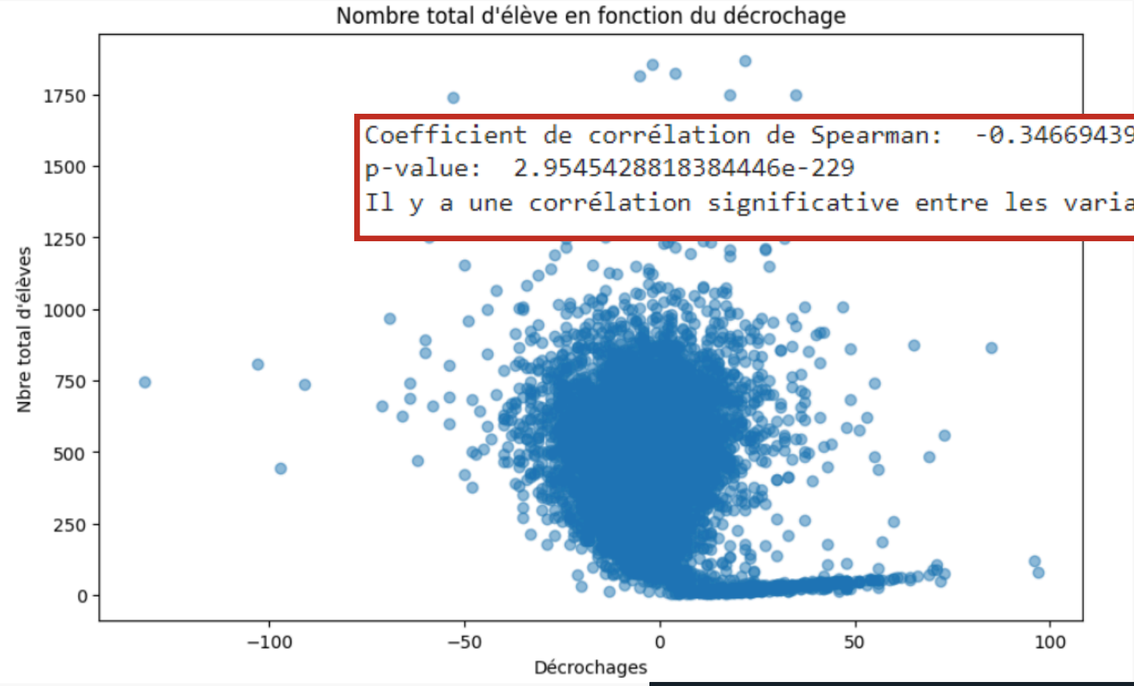
Top 10 des collèges avec les réductions les plus élevées :

	Numéro du collège	decrochage
5253	0690580F	-132
5316	0692160Y	-103
8040	9730334A	-97
774	0132915Y	-91
6636	0820683X	-71
3002	0440147W	-69
5607	0721408P	-66
4185	0593160P	-65
1924	0311266H	-64
2169	0332723F	-64

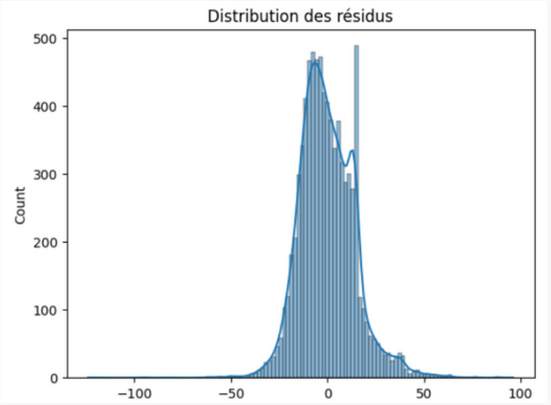
	decrochage	Nombre d'élèves total
count	8163.000000	8163.000000
mean	0.051084	415.879211
std	15.862099	260.048396
min	-132.000000	4.000000
25%	-10.000000	220.000000
50%	-2.000000	435.000000
75%	8.000000	594.000000
max	97.000000	1868.000000

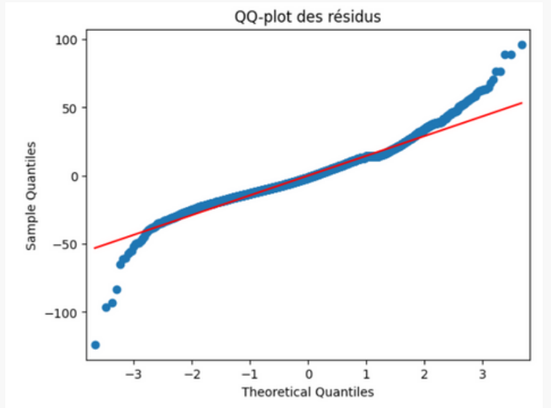
Question

Question 2 : Cette tendance a-t'elle un lien avec la taille du collège ? (volume d'étudiants)




```
Coefficient de corrélation de Spearman:  -0.3466943979982827
p-value:  2.9545428818384446e-229
Il y a une corrélation significative entre les variables et elle est de -0.35
```






```
Statistique de Shapiro-Wilk: 0.9825904369354248
p-value: 1.8559421392138536e-30
```

Question

Nous avons utilisé des tests non paramétriques

Le test de **Spearman**, aussi connu sous le nom de coefficient de corrélation de Spearman, est une mesure **non paramétrique** de la corrélation entre deux variables. Il évalue la force et la direction de l'association monotone entre les variables.

Le test de **Kruskal-Wallis** est un test **non paramétrique** utilisé pour déterminer s'il y a des **différences significatives** entre les médianes de **trois groupes ou plus**. C'est une extension du test de Mann-Whitney pour plus de deux groupes. -> ANOVA

Question

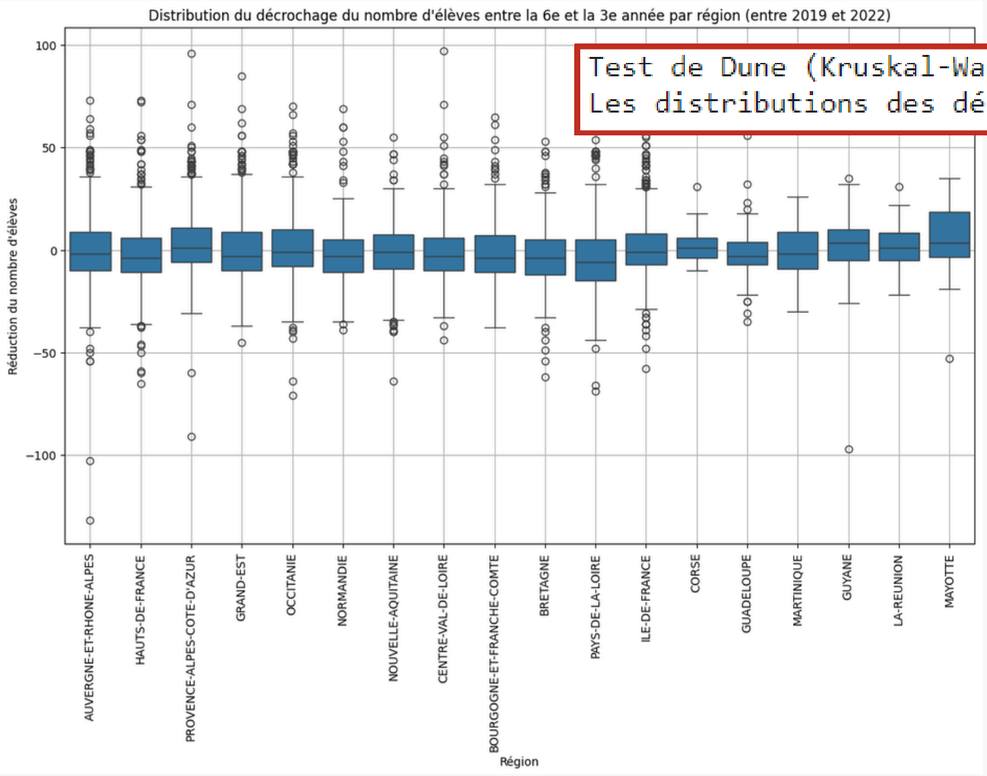
Question 3 : Est-ce la même conclusion pour les garçons que pour les filles ?

```
Corrélation de Spearman pour les garçons : coef=0.03, p_value=0.004
Corrélation de Spearman pour les filles : coef=0.03, p_value=0.005
```

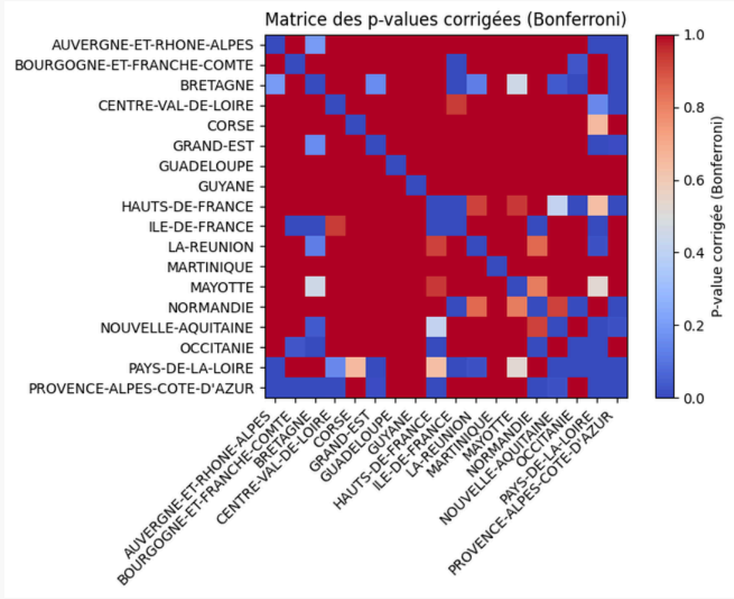
```
Statistique du test de Kruskal : 97.8153149966901
p-value : 4.5928160567642216e-23
On rejette l'hypothèse nulle : il y a une différence significative entre les décrochages des filles et des garçons.
```

Question

Question 4 : Est-ce que le constat change en fonction de la région ?

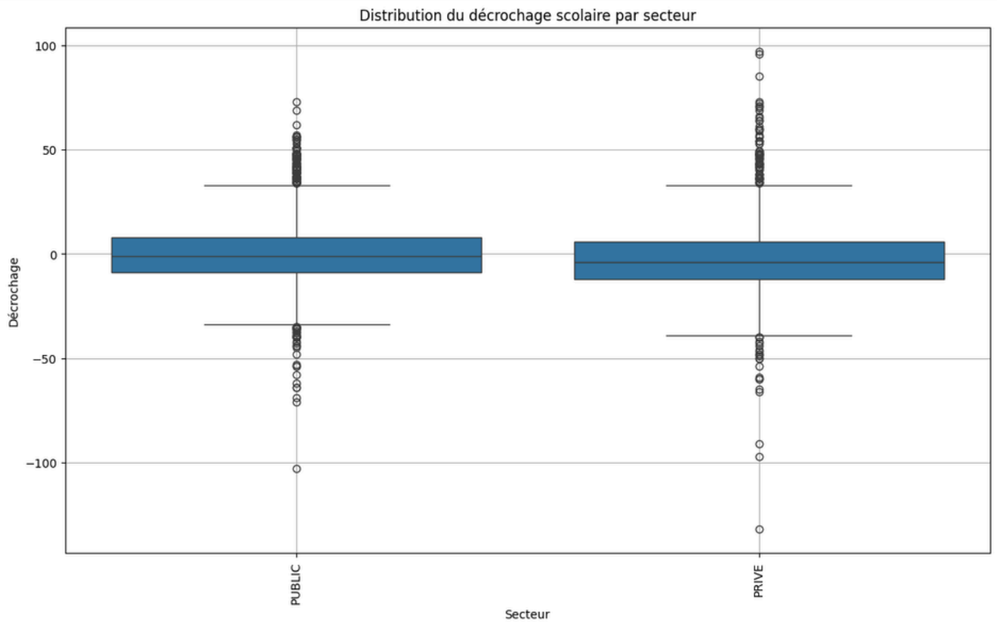


Test de Dune (Kruskal-Wallis) - Statistique : 147.4216337126249, p-value: 7.815786607433378e-23
Les distributions des décrochages par région sont significativement différentes



Question

Question 5 : Y-a-t'il un lien avec le fait d'être en collège public ou privé ?




```
Test de Dune (Kruskal-Wallis) - Statistique : 48.55201383310481, p-value: 3.2164386215511064e-12
Les distributions des décrochages par secteur sont significativement différentes, donc il y a un lien entre le secteur et le décrochage
```