

# Projet d'application Data Science

Durée de retour à l'emploi :  
les déterminants endogènes et exogènes



Emmanuelle MEUNIER

DESU « Data science pour les professionnels » 2022

30 juin 2022

Diplôme d'Études Supérieures Universitaires  
**DATA SCIENCE**  
**POUR LES PROFESSIONNELS**



Faculté d'Économie et de Gestion d'Aix-Marseille Université



# Sommaire

- 01 Problématique
- 02 Constitution du jeu de données
- 03 Statistiques descriptives
- 04 Statistiques prédictives
- 05 Résultats

# 01 Problématique

## Organisation

Estimer le temps de retour à l'emploi de salariés bénéficiant d'un « congé de reclassement » suite à un licenciement économique :

- pour optimiser les ressources (CR, cellule de reclassement, ...)
- à partir des datas internes sur les salariés / leur entreprise, croisées avec des indicateurs de vitalité du marché du travail local

Vérifier les hypothèses quant aux déterminants de l'employabilité et du repositionnement professionnel (hors facteurs psychologiques) :

- âge et ancienneté : expérience versus adaptabilité (ni trop ni trop peu)
- genre : métiers masculins / féminins, freins sociaux à l'emploi des femmes
- qualification, CSP : à mettre en perspective avec la typologie du tissu économique local
- métier : le casse-tête des nomenclatures métiers
- taille de l'entreprise d'origine : l'effet notoriété
- secteur d'activité : en émergence / en décroissance, gisement d'emplois, ...
- territoire : taux de chômage, indice de concentration d'emploi, ... et + si affinités

2 notebooks :

1. Phase exploratoire (constitution de la BDD)
2. Analyse descriptive & prédictive

Consultables sur le Github

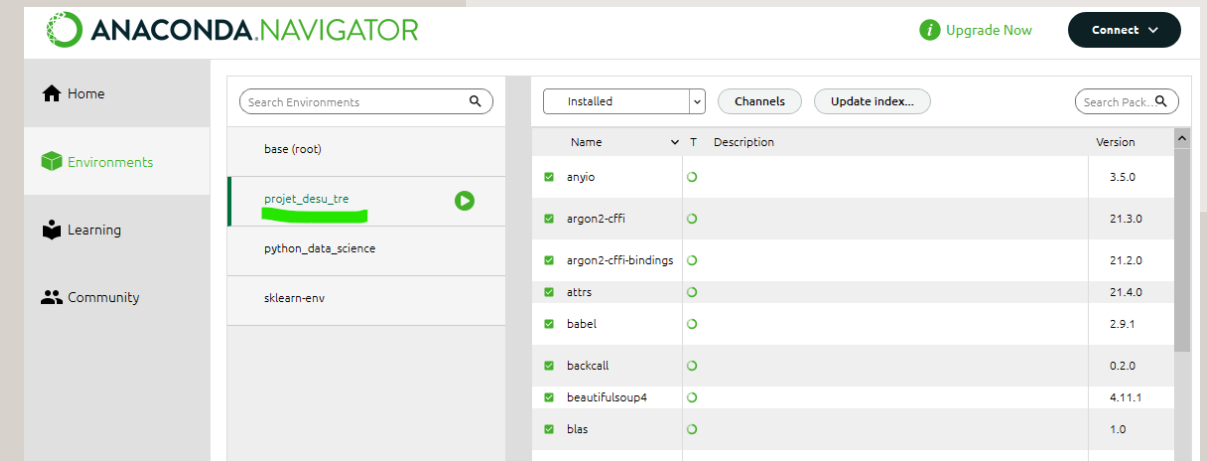
➔ [https://github.com/EmmaManou/projet\\_desu\\_tre](https://github.com/EmmaManou/projet_desu_tre)

Arborescence cookiecutter.

Organisation du dossier "data" :

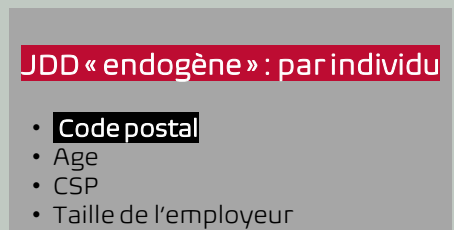
- BDD endogène (répertoire "raw")
- JDD exogènes (répertoire "external")
- BDD consolidée (répertoire "processed")

Environnement virtuel : `projet_desu_tre`

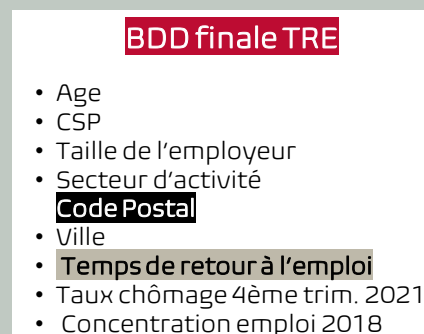


# 02 Constitution du jeu de données

## Traitement



49 999 lignes



49 999 lignes

1. Import en formatant les types (clé en format string) - dtypes
2. Suppression des colonnes inutiles
3. Identification et suppression des doublons - duplicated.sum
4. Nombre de valeurs totales (count) et valeurs différentes - nunique
5. Vérification du nombre de valeurs manquantes - isna.sum
6. Suppression des lignes avec valeurs manquantes (outremer) - dropna
7. JOIN par la clé de jointure code commune ou code zone d'emploi
8. Au besoin complétion des valeurs manquantes (arrondissements de Marseille, Lyon, Paris) - .loc
9. Suppression des colonnes inutiles et colonnes de jointure
10. Export de la BDD en format CSV - .to\_csv

	id	genre	prenom	nom	age	csp	nbr_de_salaries	ville	duree_rae	ze	chomage_2021T4	concentr_emp2018
0	1	h	Daniel	VOL	39	Employé	208	Reinhardsmunster	351	Sarrebouurg	5.4	12.7
1	2	f	Valentine	GAUGET PETIT	32	Cadre supérieur	324	Villefranche-de-Conflent	215	Perpignan	12.1	179.5
2	3	f	Angèle	BELAOUAZZA	21	Cadre supérieur	196	Saint-Gervais-les-Bains	270	Le Mont Blanc	4.1	68.9
3	4	f	Jeanine	HERMOSA GUIRADO	39	Cadre	538	La Verpillière	166	Bourgoin-Jallieu	6.3	91.1
4	5	h	Amaury	HALBERT	35	Ouvrier	930	Montpellier	45	Montpellier	9.8	151.3
...	...	...	...	...	...	...	...	...	...	...	...	...
89909	49995	h	Ghislain	ASATEKIN	29	Technicien Maîtrise	2965	Toulouse	0	Toulouse	7.4	152.4
89915	49996	h	Julien	PAUL DUBAIL	33	Cadre	452	Marseille 13e Arrondissement	168	Marseille	9.6	112.0
89916	49997	h	Hercule	UKOBIZABA	30	Cadre	5010	Nérès-les-Bains	55	Montluçon	8.8	101.0
89917	49998	f	Vienne	FLAUJAT	54	Cadre supérieur	586	Plateau d'Hauteville	33	Belley	5.5	118.2
89918	49999	f	Edwige	NKU	41	Technicien Maîtrise	426	Saint-Priest-sous-Aixe	209	Limoges	6.7	20.4

49999 rows × 12 columns

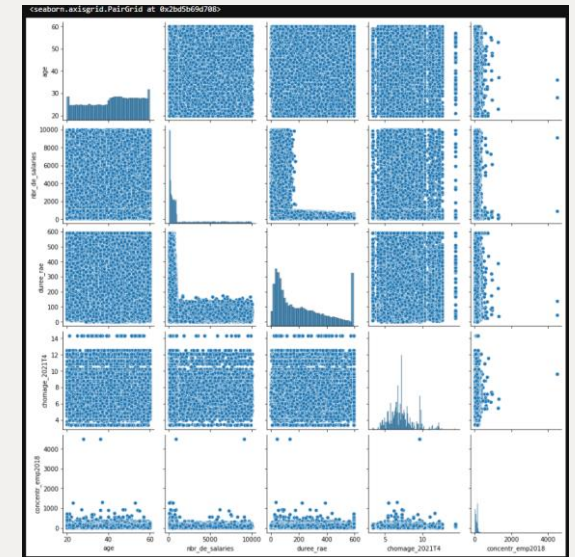
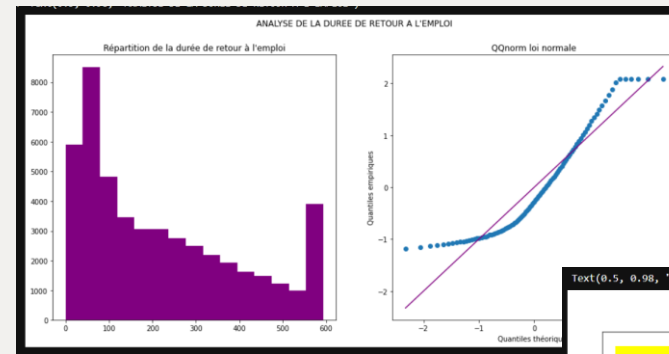
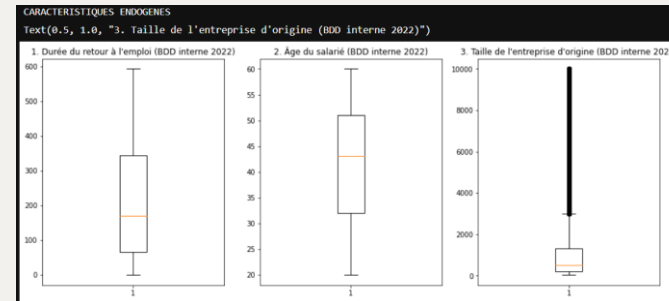
# 03 Statistiques descriptives

## Première overview

- résumé statistique, exploration des données
- identification et explication ou nettoyage des valeurs incohérentes (valeurs aberrantes, durée de reclassement < 0, nombre de salariés = 0, ...)

## Analyse des différentes variables

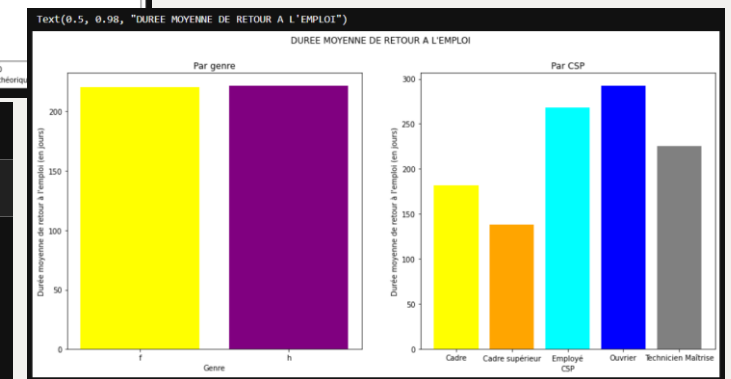
- Visualisation des variables quantitatives : boxplots
- Visualisation des corrélations : traitement différentiel entre variables quantitatives et catégorielles (pairplots)
- Représentation graphique de la variable d'intérêt : hist, QQNorm
- Variables catégorielles
- Coefficient de corrélation de Pearson



### 3.6 Corrélation de Pearson entre les différentes variables

```
[31]: # affichage des valeurs de corrélation de Pearson
tre_df.corr()
```

[31]:	age	nbr_de_salaries	duree_rae	chomage_2021T4	concentr_emp2018
age	1.000000	-0.000155	-0.036889	-0.002796	-0.006511
nbr_de_salaries	-0.000155	1.000000	-0.529443	-0.004887	-0.000666
duree_rae	-0.036889	-0.529443	1.000000	-0.014208	-0.090398
chomage_2021T4	-0.002796	-0.004887	-0.014208	1.000000	0.001587
concentr_emp2018	-0.006511	-0.000666	-0.090398	0.001587	1.000000



# 04 Statistiques prédictives

## Préparation du dataframe

- Suppression des colonnes inutiles
- One Hot Encoding des variables catégorielles

## Régression linéaire

- Constitution des échantillons LEARN & TEST
- Entraînement du modèle
- Vérification du modèle : caractère gaussien, homoscedasticité, ...
- Test du modèle : indicateurs de performance (R2, RMSE, ...)

OLS Regression Results

Dep. Variable:

duree\_rae

R-squared:

0.389

Model:

OLS

Adj. R-squared:

0.389

Method:

Least Squares

F-statistic:

2244.

Date:

Sun, 26 Jun 2022

Prob (F-statistic):

0.00

Time:

14:30:20

Log-Likelihood:

-2.0188e+05

No. Observations:

31758

AIC:

4.038e+05

Df Residuals:

31748

BIC:

4.039e+05

Df Model:

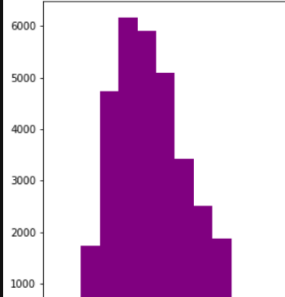
9

Covariance Type:

nonrobust

	coef	std err	t	P> t	[0.025	0.975]
age	-0.6361	0.069	-9.178	0.000	-0.772	-0.500
nbr_de_salaries	-0.0362	0.000	-120.882	0.000	-0.037	-0.036
chomage_2021T4	-2.1132	0.473	-4.467	0.000	-3.040	-1.186
concentr_emp2018	-0.2261	0.011	-19.869			
genre_f	102.2897	1.596	64.075			
genre_h	103.8086	1.593	65.179			
csp_Cadre	2.6998	1.664	1.622			
csp_Cadre supérieur	-41.8346	1.652	-25.322			
csp_Employé	84.0179	1.668	50.362			
csp_Ouvrier	114.1871	1.660	68.779			
csp_Technicien Maîtrise	47.0280	1.662	28.297			
const	206.0983	2.778	74.186			

Caractère gaussien de



A histogram of residuals, likely from the OLS regression, showing a distribution that is roughly symmetric and bell-shaped, consistent with the Gaussian assumption. The x-axis is labeled 'Caractère gaussien de' and the y-axis shows frequency counts from 0 to 6000.

Omnibus:

1241.151

Durbin-Watson:

Prob(Omnibus):

0.000

Jarque-Bera (JB):

Skew:

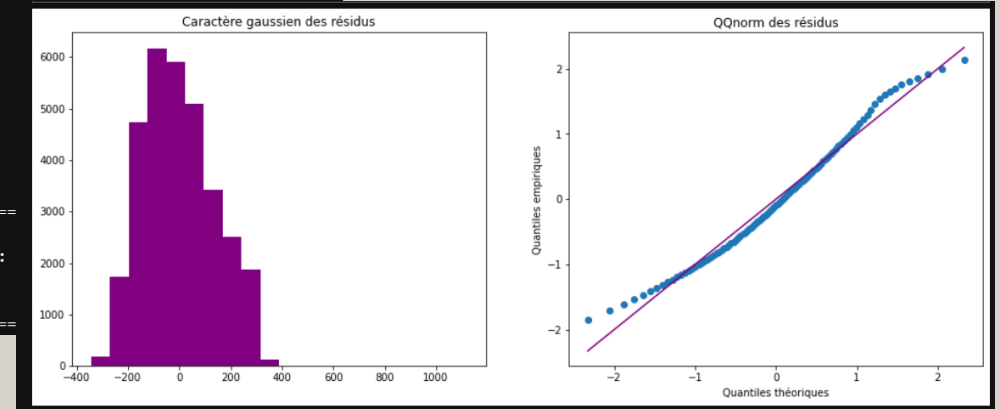
0.325

Prob(JB):

Kurtosis:

2.471

Cond. No.



# 05 Résultats

Une infirmation des hypothèses à mettre en regard avec le caractère aléatoire du jeu de données :

- absence de corrélation entre variables dès la phase descriptive
- logiquement, la régression linéaire n'a pas permis de proposer un modèle performant
- pas de tests d'autres modèles qui auraient pu être plus performants

Un travail à réitérer sur un « vrai » jeu de données :

- confirmer ou infirmer, grâce à l'analyse descriptive, les hypothèses issues de la littérature et de l'expérience
- disposer de plus de variables explicatives à exploiter
- explorer la possibilité de traiter la variable catégorielle "métier" (traitement préalable par NLP ?)
- creuser les indicateurs de dynamisme du marché du travail (complémentarité, redondance, ...)

Réentraîner avec le modèle linéaire ... et plus si affinités :

- en proposant une validation croisée sur la partie TRAIN (K-fold) et en minimisant la fonction de coût (Descente de Gradient)
- en approfondissant l'analyse des indicateurs lors de la phase de validation ( $R^2$ , RMSE, coefficients, Pvalue, ...)
- en optimisant les hyperparamètres...
- en testant d'autres méthodes de ML pour voir si elles sont plus performantes (forêt aléatoire, ...)

Construire le template adéquat pour les fonctionnels