Année universitaire : 2022 – 2023

PROJET PYTHON

Réalisé par
Eunice KOFFI
Gaoussou DIAKITE

Professeur
Hager Oueslati

Notre application : Mon logement Nexity · Streamlit

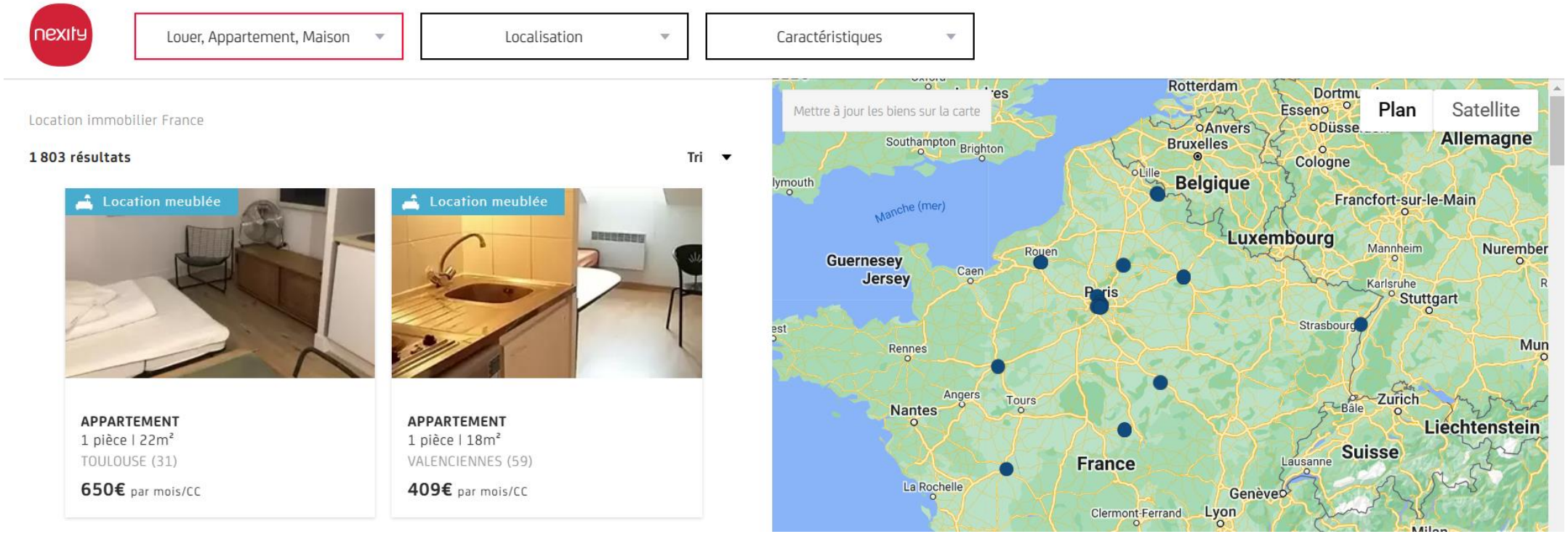
SOMMAIRE

1. Scraping
2. Features engineering
3. Analyse exploratoire
4. Machine Learning

1. Scrapping (1/2)

Nexity est un site qui propose des offres d'achat, de location ou de vente de biens immobiliers (appartements, maisons, terrains, immeubles, ...) principalement en France.

Lien du site : Location immobilier France | Nexity



1. Scrapping (2/2)

- Informations disponibles au 23/01/2023
- Bibliothèques utilisées : BeautifulSoup
- Données extraites : Toutes les informations disponibles pour chaque logement
- Résultats : 21 colonnes et 1032 observations

Un aperçu de la base de données

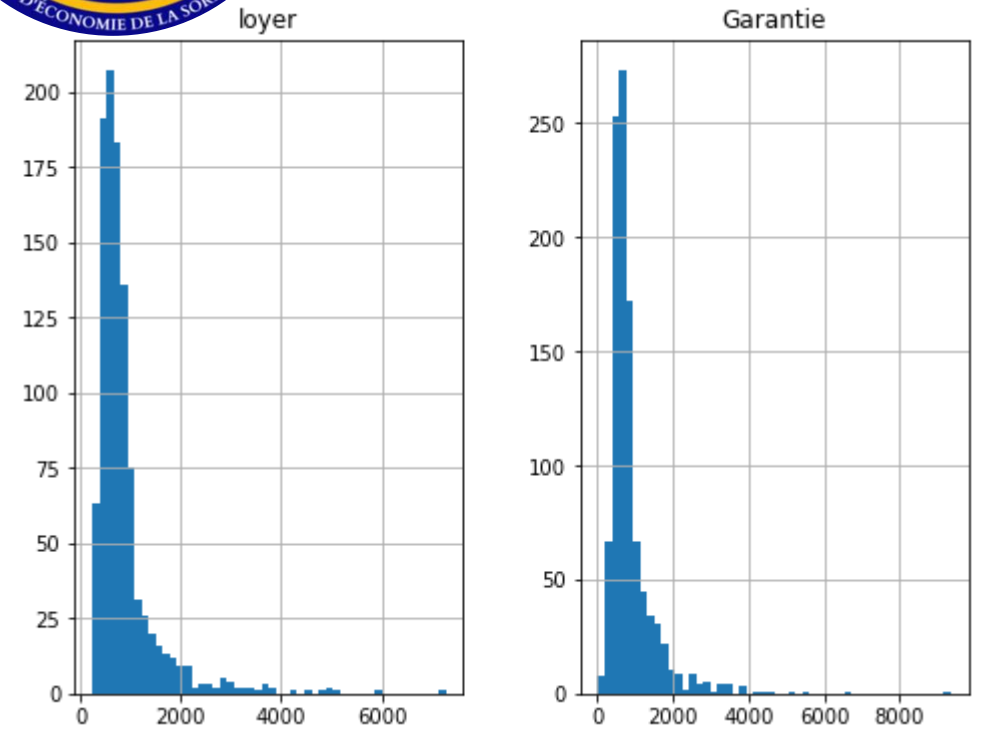
Types	Titre	Logement	Code_postal	Departement	Superficie	Essentiel	Amenagement	Prix_loyer	
location meublée	Appartement à louer :...	- Appartement studio ...	31000	Toulouse	22 m²	Type de BienApparte...	Pièce(s)1WC séparéOui	650 € CC/mois	Plein
location meublée	Appartement à louer :...	- Appartement studio ...	51100	Reims	27 m²	Type de BienApparte...	Pièce(s)1WC séparéOui	450 € CC/mois	hype
location meublée	Appartement à louer :...	- Appartement studio ...	59300	Valenciennes	18 m²	Type de BienApparte...	Pièce(s)1WC séparéOui	409 € CC/mois	
location meublée	Appartement étudiant...	NA	NA	84000	NA	Type de BienApparte...	Pièce(s)1WC séparéOui	NA	AVIG
location meublée	Appartement à louer :...	- Appartement studio ...	33600	Pessac	21 m²	Type de BienApparte...	Pièce(s)1WC séparéOui	632 € CC/mois	PES
location meublée	Appartement à louer :...	- Appartement 3 pièces	33400	Talence	69 m²	Type de BienApparte...	Pièce(s)3WC séparéOui	1 155 € CC/mois	Tale
location meublée	Appartement à louer :...	- Appartement studio ...	51100	Reims	24 m²	Type de BienApparte...	Pièce(s)1WC séparéOui	502 € CC/mois	Hyp
location meublée	Appartement à louer :...	- Appartement studio ...	51100	Reims	18 m²	Type de BienApparte...	Pièce(s)1WC1WC sé...	402 € CC/mois	Rés
location meublée	Appartement à louer :...	- Appartement 4 pièces	38000	Grenoble	115 m²	Type de BienApparte...	Pièce(s)4Chambre(s)...	1 780 € CC/mois	EX
location meublée	Appartement à louer :...	- Appartement studio ...	89000	Auxerre	17 m²	Type de BienApparte...	Pièce(s)1WC séparéOui	354 € CC/mois	Dans
location meublée	Appartement étudiant...	NA	NA	59564	NA	Type de BienApparte...	Pièce(s)1WC séparéOui	NA	L'idé

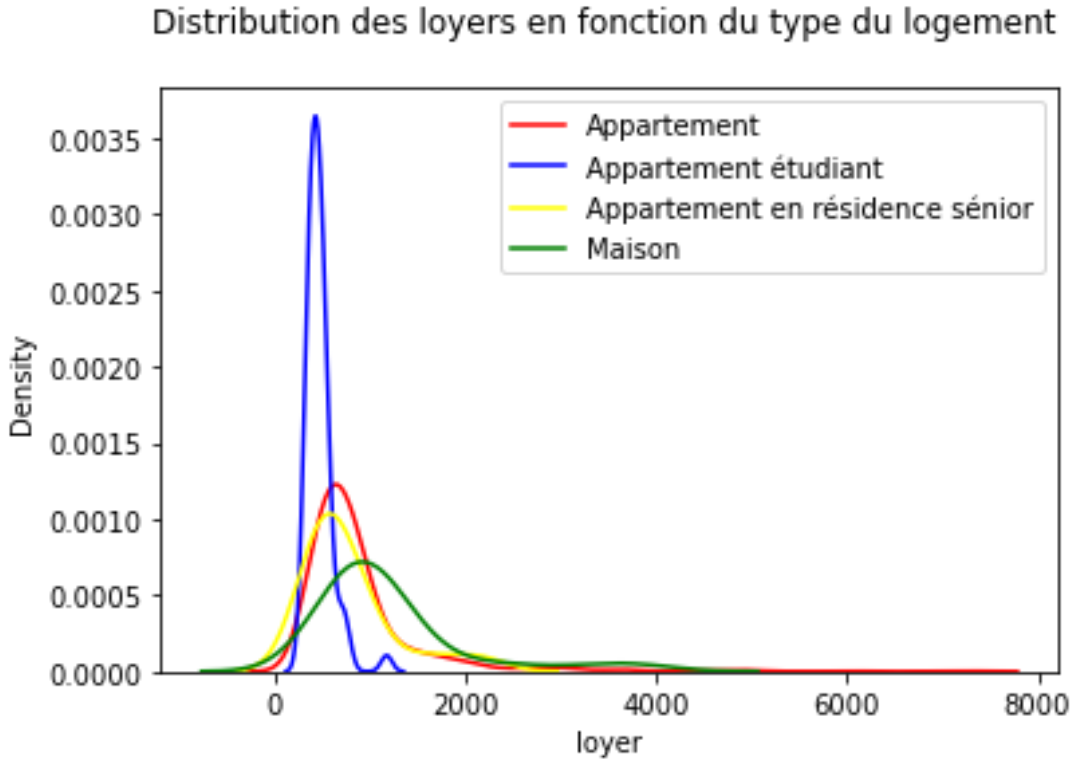
2. Features engineering

- Un constat
  - Des informations répétitives dans plusieurs variables (loyer, charge, …)
  - Des informations indisponibles dans certaines variables et présentes dans d'autres
- Sélection / Suppression de variables
  - Sélection des variables contenant des informations complètes sans valeurs manquantes (ex: la colonne titre contient à la fois le type, la superficie, le nombre de pièces et ce, de manière complète pour toutes les observations)
  - Suppression des variables initiales incomplètes qui pourraient être reconstruites (ex : loyer et charges)
- Traitement
  - Utilisation de regex
  - Split de colonnes
  - Conditions

- Résultats : 22 colonnes et 1032 observations

3. Exploration des données (1/4)



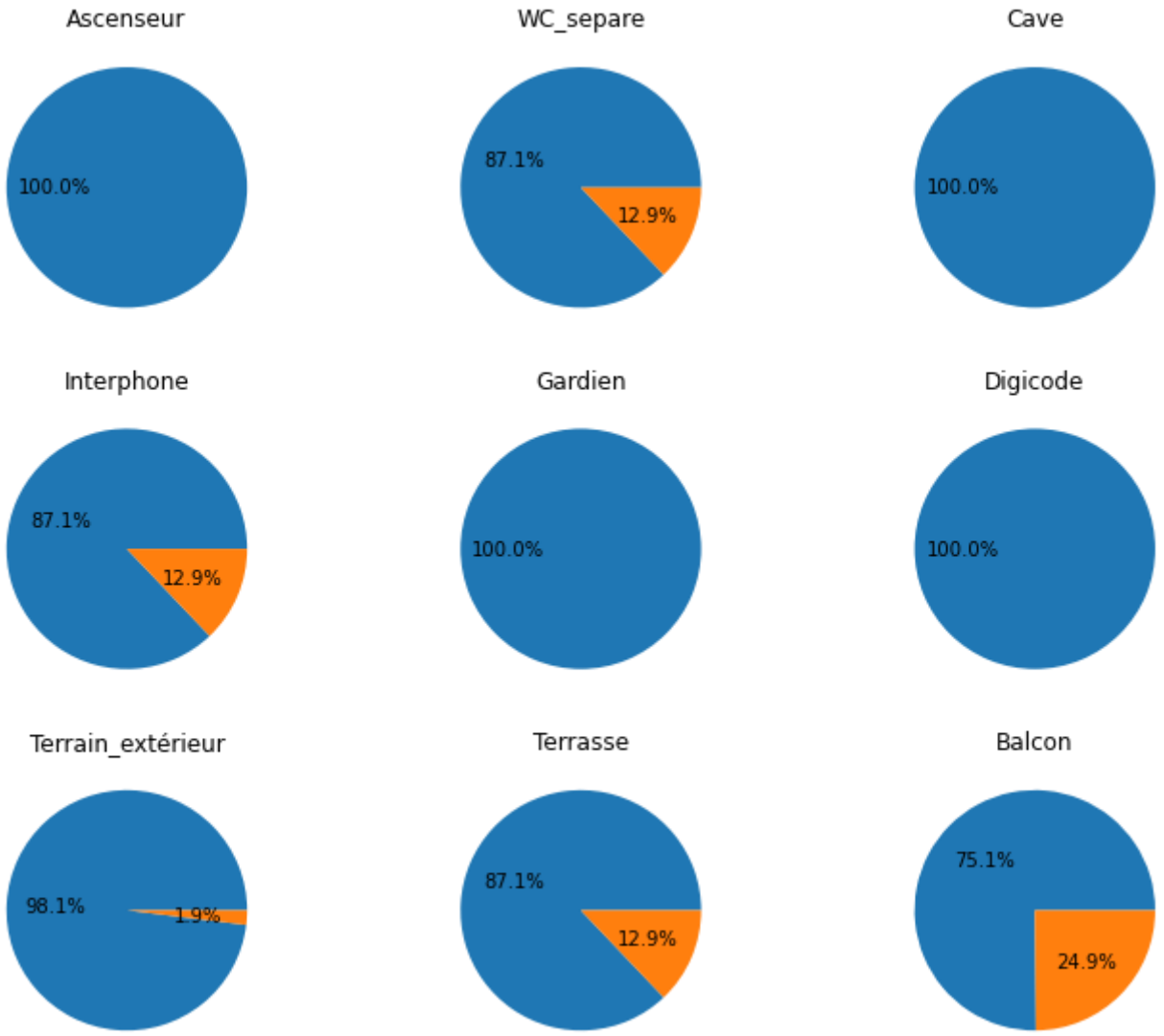


- On a une distribution asymétriques des loyers dans notre base de données. La plupart des loyers sont inférieurs à 2000 €
- On remarque une distribution des garanties similaire à celle des loyers. Ce qui est normal car la garantie correspond le plus souvent à une ou deux mensualité du loyer.

- En observant la décomposition de ces loyers en fonction du type de logement on se rend compte les appartements ont les loyers les plus élevés.
- Les appartements étudiants ont les loyers les plus faibles.

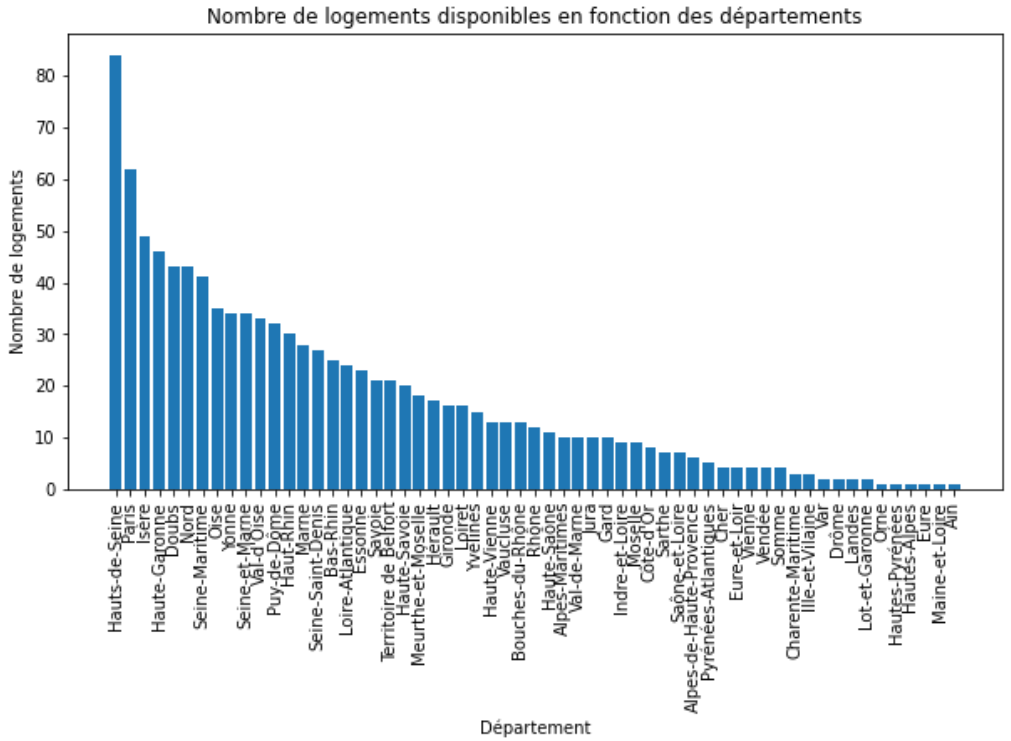
3. Exploration des données (2/4)

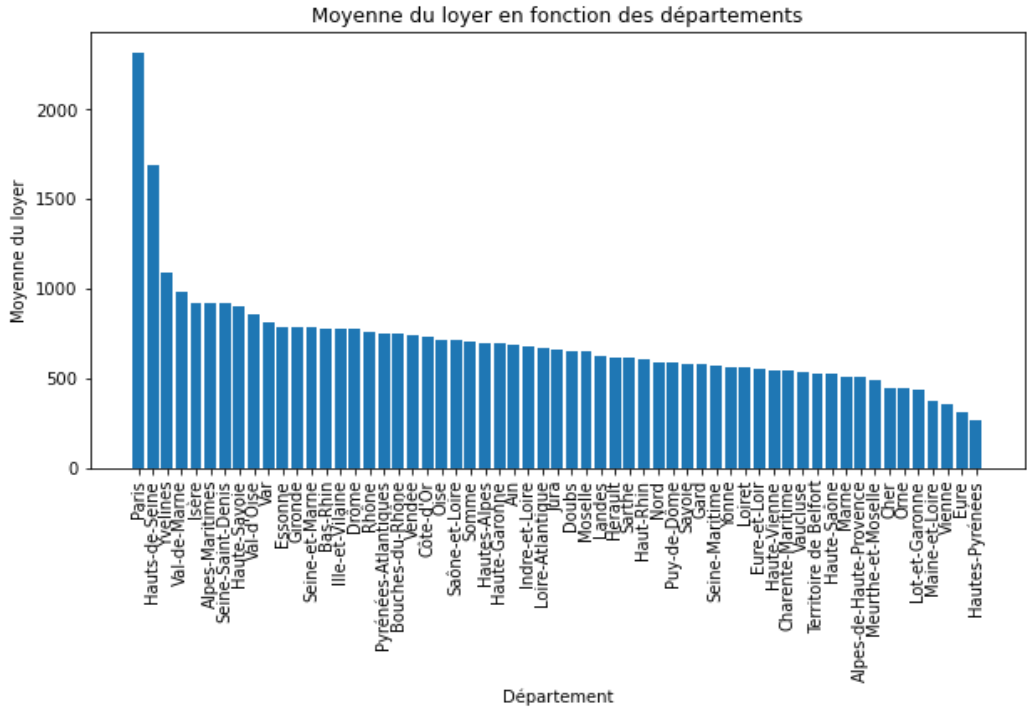
- Nous avons plusieurs variables unimodales.
- Tous nos logements sont dotés d'ascenseur, de WC séparé, Cave, Gardien, Digicode, Terrain extérieur
- La majorité ont une terrasse, un balcon et un interphone



3. Exploration des données (3/4)

- Les départements qui comptabilisent le plus de logements sont : Hauts-de-Seine, Paris, Isère, Haute-Garonne
- Cependant les départements qui ont les loyers les plus élevés sont Paris, Hauts-de-Seine, Yveline et Val-de-Marne





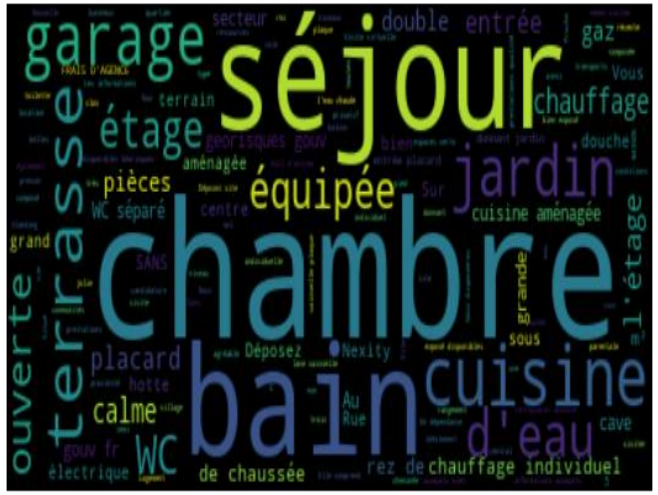
3. Exploration des données (4/4)

On affiche pour chaque type de logement, un nuage de mots formé à partir des descriptions des logements.

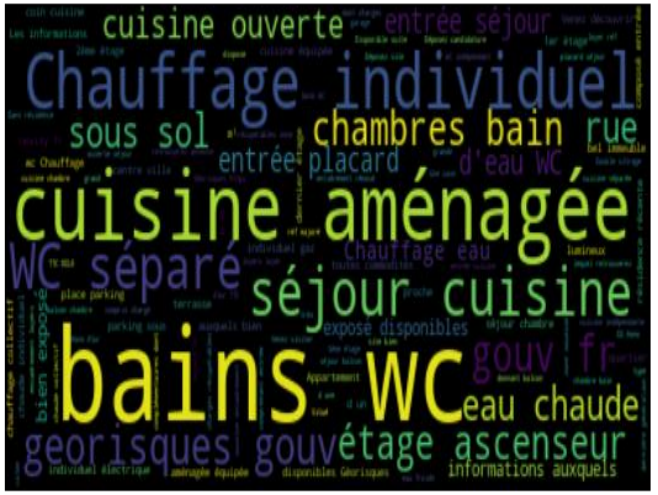
On remarque, alors, des caractéristiques déterminantes différentes selon le type de logement.

- Pour les maisons les caractéristiques mises en avant sont : Le séjour, la cuisine, la terrasse, le garage, la cuisine
- Pour les appartements, ce sont : Le chauffage, la salle de bain, les chambres, …
- Pour les résidences sénior : La terrasse, les services, la possibilité de visite, …
- Pour les résidences étudiantes : le fait que l'appartement soit meublé, la proximité avec les commerce et la gare, la présence de laverie, …

Maison

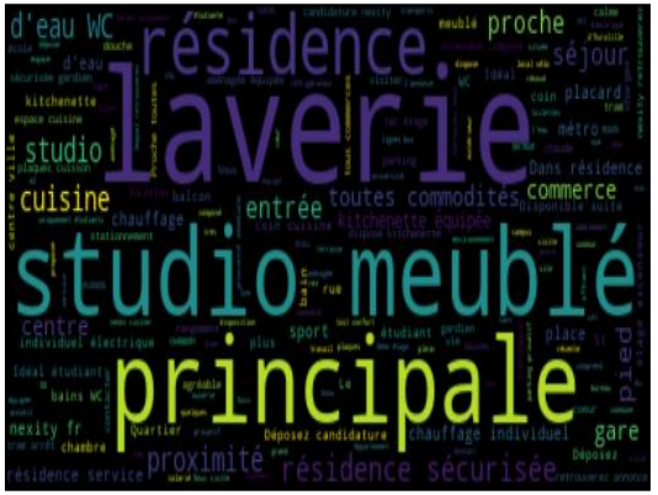


Appartement



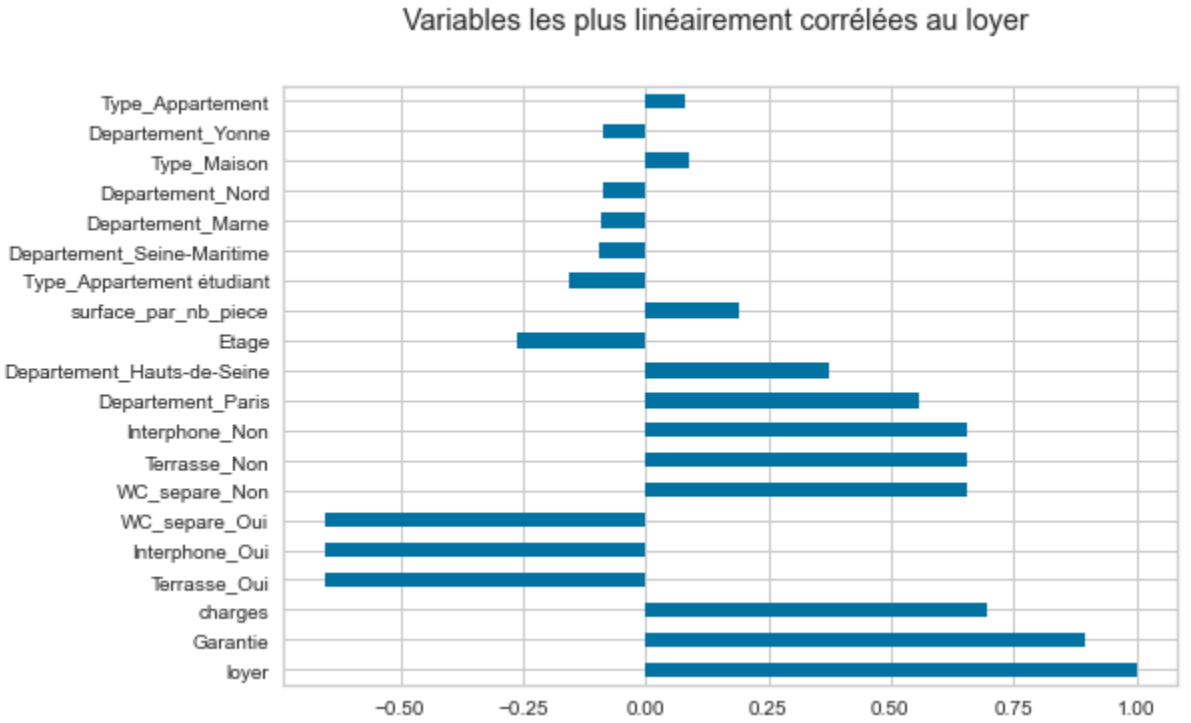
Sénior

Etudiant



4. Machine Learning (1/4)

- Comme nous pouvons le voir dans la matrice de corrélation ci-dessus, il existe deux variables qui montrent une forte multicolinéarité (**r > 0,80**) : les chambres et la surface. Cela a du sens puisque plus une maison est grande, plus elle aura généralement de chambres.
- Une façon de résoudre ce problème est de combiner les deux variables en une seule, par exemple en créant une nouvelle variable pour **la surface par chambre**.



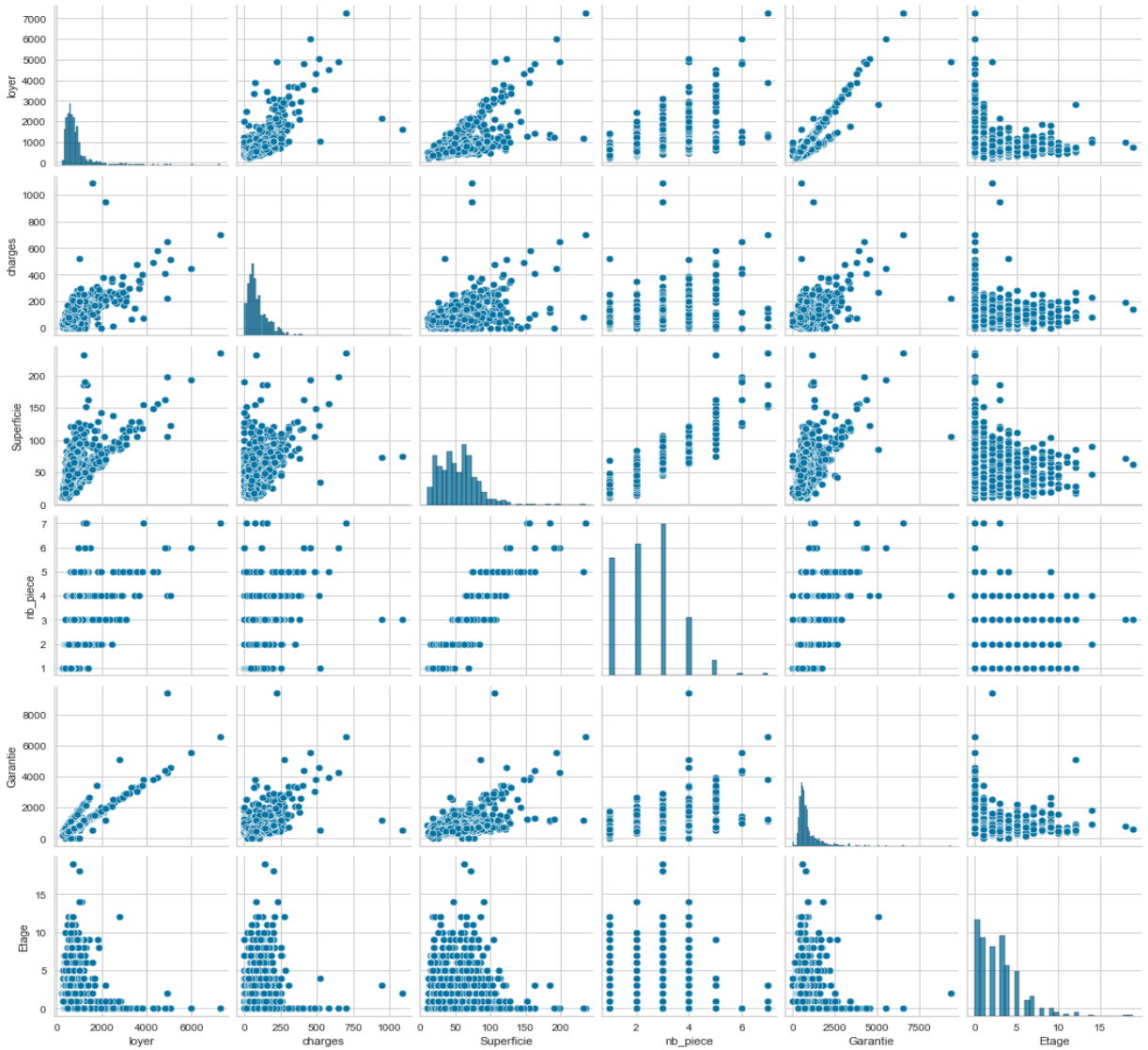
Dans une conclusion quelque peu similaire à la surface, plus une maison a de chambres, plus le loyer devrait être élevé. Cependant, nous constatons qu'un autre facteur entre en ligne de compte car il existe des appartements de 5 pièces pour lesquels le loyer est plus faible que d'autres qui ont un plus petit nombre de pièces.

4. Machine Learning (2/4)

- Préparation des données
  - Analyse des corrélations
  - Croisement de variables
  - Dichotomisation de nos catégorielles

- Le graphique nous affiche plusieurs représentations de nos données dans un espace de deux dimensions en utilisant plusieurs combinaisons de nos variables. La forme du nuage de mots nous montre clairement si une corrélation linéaire existent entre les dimensions choisies. Nous pouvons de ce fait voir que certaines variables ne sont pas corrélées (étage et nombre de pièces).

- La matrice de corrélation nous apportera une information chiffrée de nos remarques.



4. Machine Learning (3/4)

Régression linéaire

Model:	OLS	Adj. R-squared:	0.558
Method:	Least Squares	F-statistic:	109.3
Date:	Sun, 29 Jan 2023	Prob (F-statistic):	1.69e-173
Time:	18:40:14	Log-Likelihood:	-7716.8
No. Observations:	1030	AIC:	1.546e+04
Df Residuals:	1017	BIC:	1.552e+04
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
surface_par_nb_piece	9.2561	2.670	3.467	0.001	4.017	14.495
Type_Appartement	-28.9293	37.755	-0.766	0.444	-103.016	45.158
Type_Appartement en résidence sénior	98.1792	88.763	1.106	0.269	-76.000	272.358
Type_Appartement étudiant	-158.4497	54.335	-2.916	0.004	-265.072	-51.828
Type_Maison	421.9610	68.803	6.133	0.000	286.950	556.972
WC_separe_Non	280.3954	25.913	10.821	0.000	229.546	331.245
WC_separe_Oui	52.3658	15.460	3.387	0.001	22.029	82.702
Interphone_Non	280.3954	25.913	10.821	0.000	229.546	331.245
Interphone_Oui	52.3658	15.460	3.387	0.001	22.029	82.702
Terrain_extérieur_Non	169.3340	46.070	3.676	0.000	78.931	259.737
Terrain_extérieur_OUI	163.4272	60.749	2.690	0.007	44.219	282.636
Terrasse_Non	280.3954	25.913	10.821	0.000	229.546	331.245
Terrasse_Oui	52.3658	15.460	3.387	0.001	22.029	82.702
Balcon_Non	103.8417	21.000	4.945	0.000	62.634	145.049
Balcon_Oui	228.9195	26.202	8.737	0.000	177.504	280.335
Departement_Paris	1155.0913	80.337	14.378	0.000	997.445	1312.737
Departement_Val-d'Oise	110.4898	38.885	2.841	0.005	34.186	186.794
Departement_Hauts-de-Seine	466.9149	81.702	5.715	0.000	306.591	627.239
Departement_Isère	208.2336	65.056	3.201	0.001	80.575	335.892
Departement_Seine-Saint-Denis	237.5101	85.939	2.764	0.006	68.871	406.149
Departement_Val-d'Oise	110.4898	38.885	2.841	0.005	34.186	186.794

Omnibus:	756.589	Durbin-Watson:	1.815
Prob(Omnibus):	0.000	Jarque-Bera (JB):	29077.764
Skew:	2.914	Prob(JB):	0.00
Kurtosis:	28.369	Cond. No.	1.92e+20

- **R^2 = 0.56**

Notre modèle arrive à expliquer 56% de la variance du loyer. On a un bon ajustement de notre modèle aux variables

- **Prob F-statistic = 1.69e-173**

Au seuil de 5 %, on rejette l'hypothèse nulle selon laquelle notre modèle n'est pas globalement significatif.

- **Prob(JB) = 0.00**

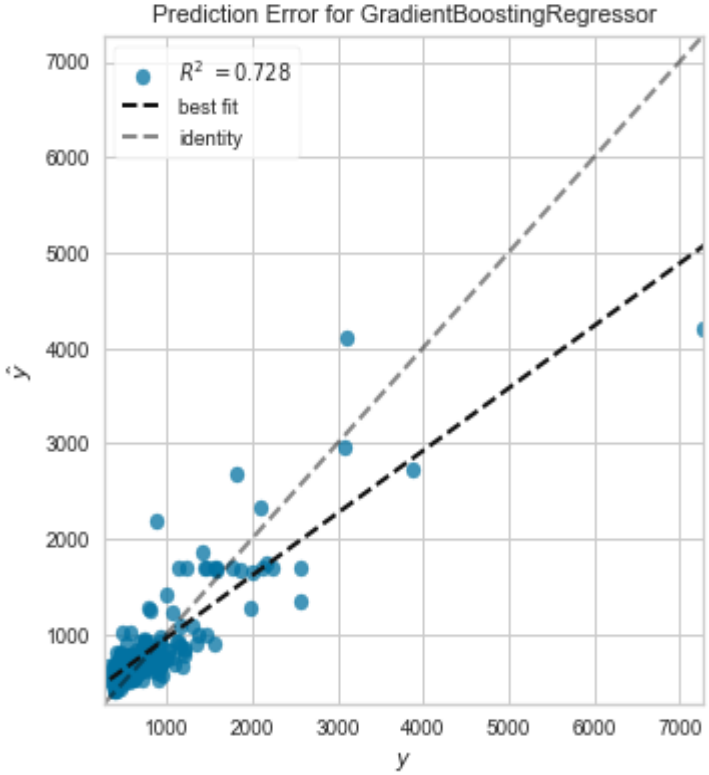
Au seuil de 5 %, on rejette l'hypothèse nulle selon laquelle nos résidus ne suivent pas une loi normale.

- **Les variables explicatives ne sont pas toutes significatives.**

La surface par piece est significative : Si la surface par pièce augmente de 1m² le loyer augmentera de 9, 25 € toutes choses égales par ailleurs.
Le fait qu'un appartement soit un appartement étudiant dimuniera le loyer de 158, 45 € en moyenne par rapport aux autres toutes choses égales par ailleurs.
Le fait un logement se situe à Paris augmentera le loyer en moyenne de 1155 € toutes choses égales par ailleurs.

4. Machine Learning (4/4)

Gradient Boosting




```
Score Test: 0.7282956149252022
MAE: 215.1918036428456
MSE: 129123.35479682879
```

Le modèle a de meilleures performances que la régression linéaire effectuée.


FALEMINDERIT
ΕΥΧΑΡΙΣΤΩ
TÄNAN
GRACIAS
DZIĘKUJĘ
TACK
GRAZIE
ありがとう
HVALA
DANKE
PALDIES
ASÂNTE
MERCI
شكرا لك
*NA GODE
ДЗЯКУЮ
DIAKUIU
ACIU
TAK
THANK YOU
DANK U WEL
धन्यवाद
СПАСИБО
DANKON
谢谢
OBRIGADO
diolch
KIITOS
TESEKKUR EDERIM
Adobe Stock | #98542148