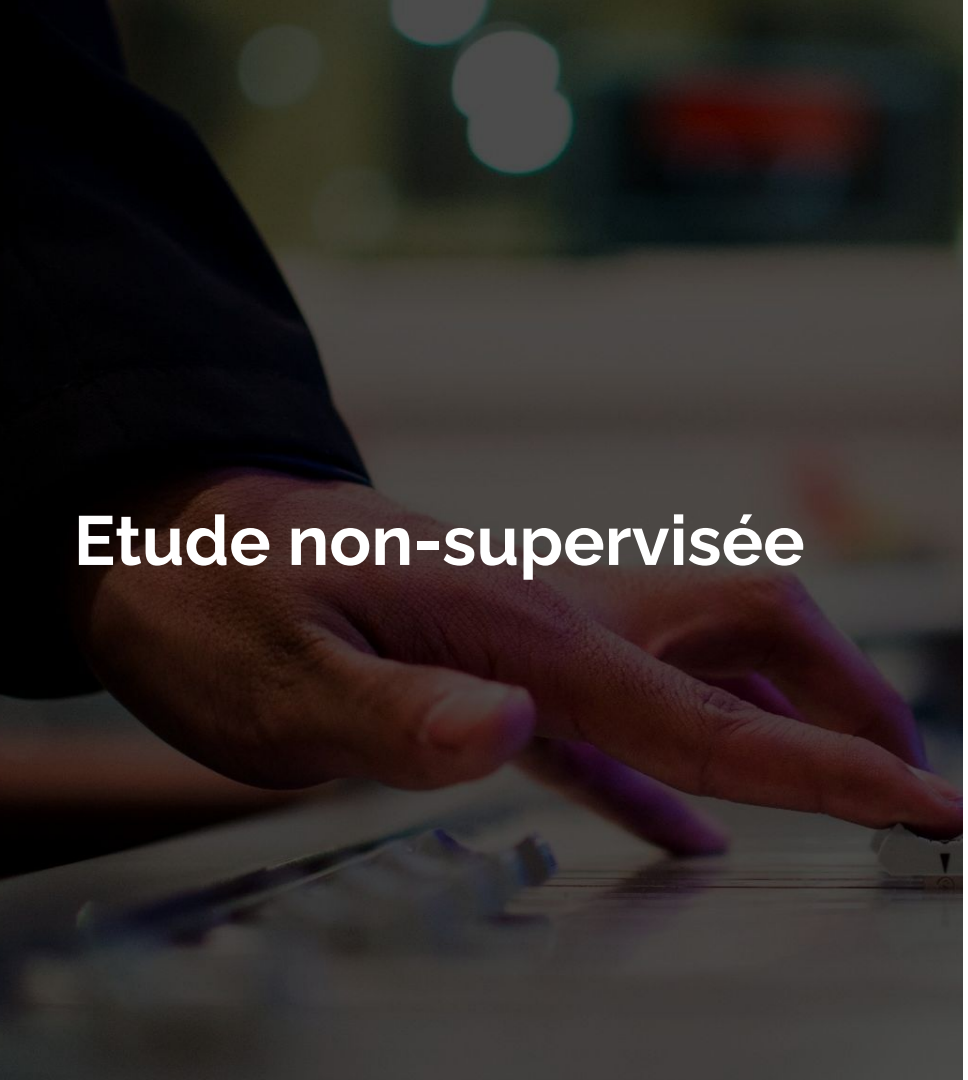
PYDS 2023 - San Francisco Rentals

Presentation du dataset

- San Francisco Rental data (2022, TidyTuesday)
- Variables : **loyer, nb chambres, nb salles de bain, surface, ville, comté, quartier et an de l'annonce**
- Après nettoyage : 260k -> 14k individus
    - Suppression des lignes avec NaN, outliers …

- Objectif : **Déterminer quels paramètres influencent les prix des loyers le plus (non-supervisé) et tester des modèles de prédiction pour les prix des loyers (supervisé)**



Etude non-supervisée

1. **Methode des K-Means**
    a. Déterminer des relations cachées entre variables par visualisation des clusters

2. **Reduction de dimension (ACP)**
    a. Déterminer quelles variables ont moins de poids sur les modeles de prediction

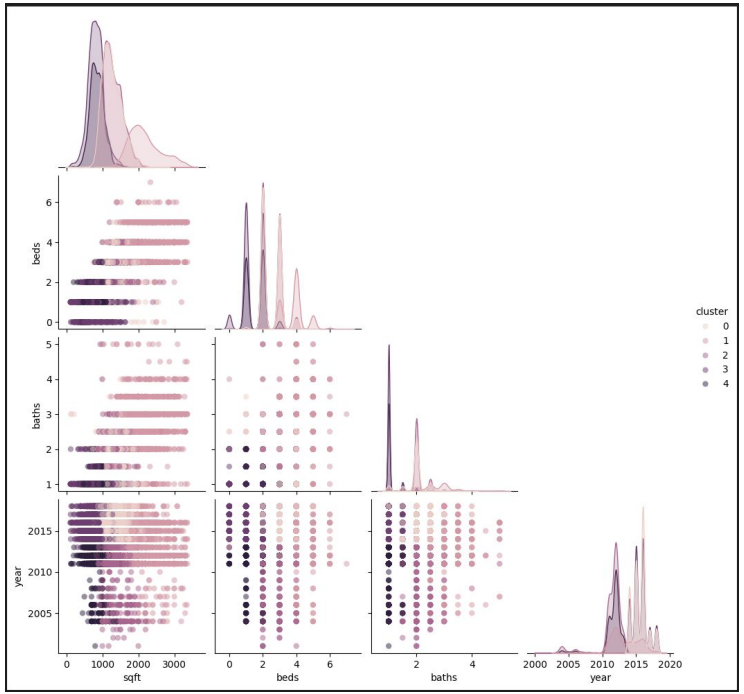
Résultats des méthodes N-S

Clustering avec K-Means :

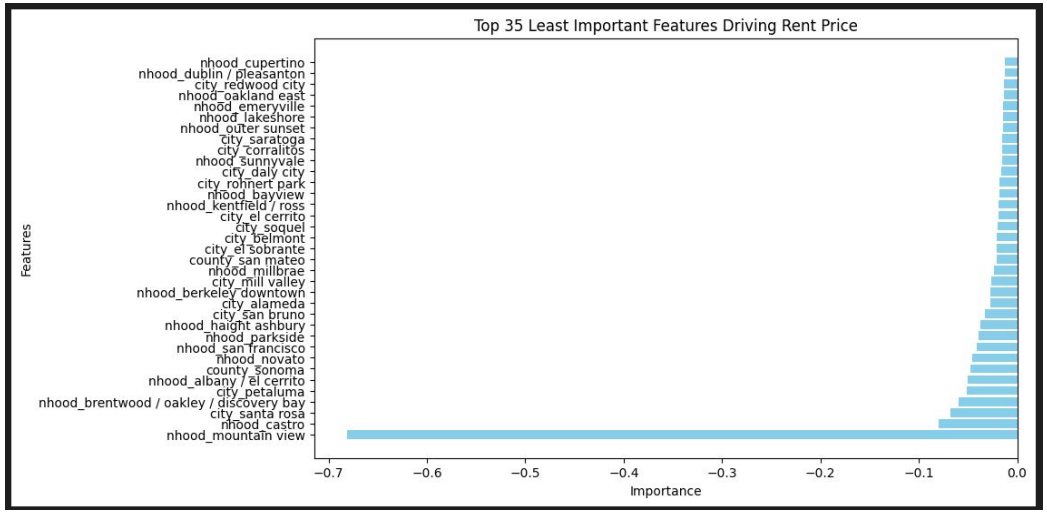
- KMeans avec 5 clusters : sqft, chambres, salles de bain, année.
- Consistance dans les schémas de distribution des clusters pour la plupart des combinaisons de caractéristiques.
- Similarités entre lits vs sqft, salles de bain vs sqft, lits vs salles de bain, salles de bain vs lits. Faible densité pour salles de bain vs lits, année vs lits et année vs salles de bain.

Réduction de Dimensionnalité avec PCA :

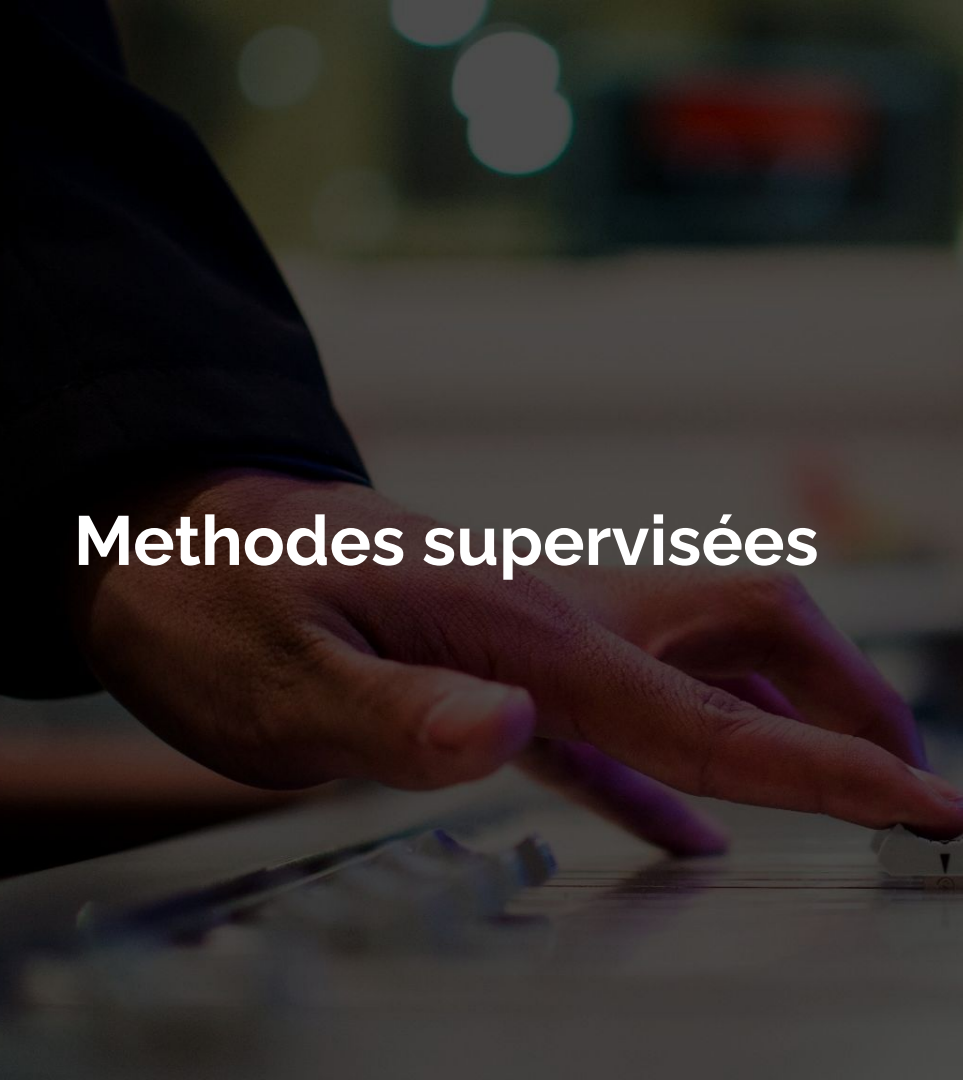
- Caractéristiques immobilières (sqft, chambres, salles de bain) dominantes pour les prix de location.
- Géographie moins influente : quartier, ville, comté.
- Attributs des biens immobiliers principaux pour les prix de location, suggérant une forte prépondérance.



Plot seaborn K-Means



35 features les moins importantes d’apres la methode de reduction de dimension



Methodes supervisées

1. **Elastic Net Regression**
    a. Avoir la flexibilite de Ridge/Lasso

2. **Random Forest Regressor**
    a. Avoir la robustesse d'un modèle à base d'arbres de décision

3. **Gradient Boosting**
    a. Avoir un modèle qui se construit de manière itérative pour donner une bonne précision finale

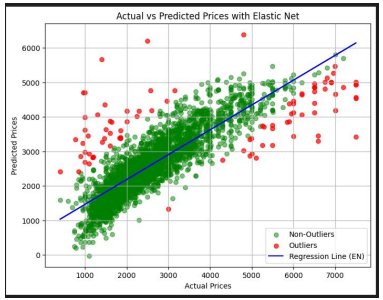
Résultats des méthodes supervisées

Elastic Net

R2 = 0.71

RMSE = 600

Prédictions globalement serrées

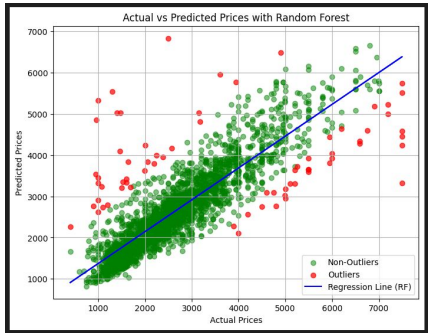


Random Forest

R2 = 0.76

RMSE = 545

Prédictions plus serrées

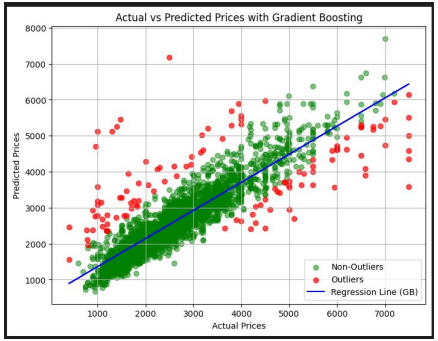


Gradient Boosting

R2 = 0.78

RMSE = 527

Prédictions les plus serrées



Comparaison des méthodes

Elastic Net :

- L1 à 0,9, se rapprochant de Lasso, a été trouvé comme idéal.
- Le modèle obtenu était précis.
- Les caractéristiques importantes diffèrent de celles des autres modèles.
- Cependant, une certaine inexactitude a été constatée sur les extrêmes des valeurs prédites.

Random Forest :

- Les estimateurs ont été fixés à 100 pour obtenir des résultats plus précis.
- Les caractéristiques importantes étaient similaires à celles de Gradient Boosting, mais différentes de celles d'Elastic Net.
- La dispersion des résidus était moindre que celle d'Elastic Net.

Gradient Boosting :

- Les estimateurs ont été fixés à 100 et ont fourni les résultats les plus précis.
- Des caractéristiques importantes similaires à Random Forest.
- La dispersion des résidus était plus faible que celle de Elastic Net, mais présente.

Cependant, une certaine inexactitude a été constatée sur les extrêmes des valeurs prédites pour les trois modèles!

Conclusion