

# Projet régression linéaire

Base de données immobilière

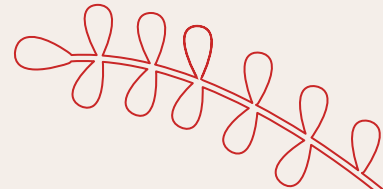
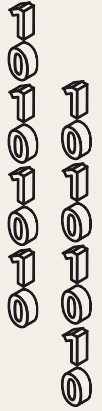




01

# Contexte

Bloc d'habitations en Californie

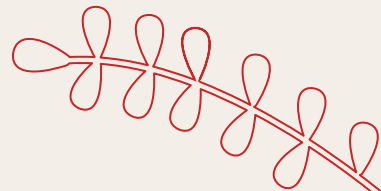




# 02 Traitement des données



Remarques et hypothèses



# 3 problèmes remarqués très tôt



## ● Une colonne sans nom

*Après vérification, colonne ID*

Inutile pour le modèle

## ● Des doublons en coordonnées

*Problématique car les valeurs des autres colonnes sont différentes*

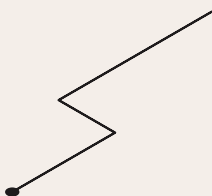
*À traiter ou non ?*

## ● Des valeurs manquants

*Environ 1% de la colonne total\_bedrooms*

Plusieurs manières de la traiter

Meilleurs résultats : KNN imputer



# Autres points importants

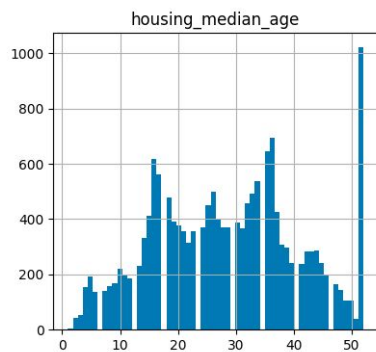
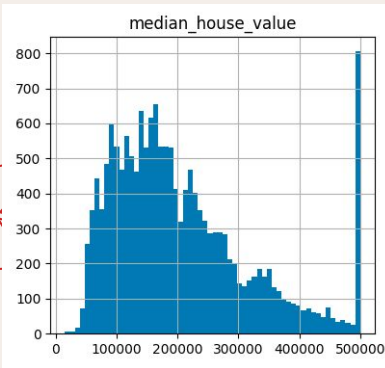


## Ocean proximity

### Colonne qualitative

À encoder.

Variable nominale ou ordinale ?

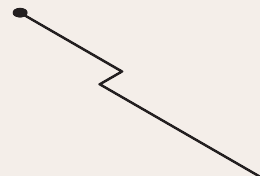


## Outliers

### House\_age & value

Évoque des valeurs bloquées à un certain maximum.

Bins pour les âges



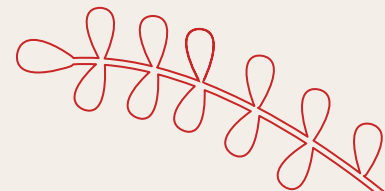
3

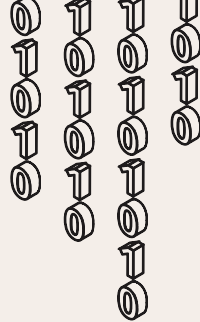
# Meilleurs résultats

Modèle linear regression



01010101  
01010101





# Les 3 transformations les plus pertinentes



## Imputation KNN

On évite de perdre une colonne utile



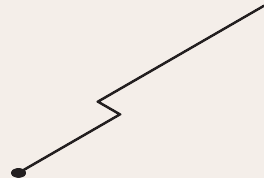
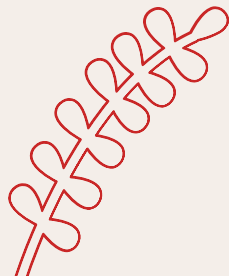
## Bins et encodage

Permet de gérer des outliers sans les retirer, et d'utiliser les variables qualitatives



## Transformation log Et colonne outliers

Permet de normaliser les données sans perdre l'information sur les outliers



# Resultats du modèle

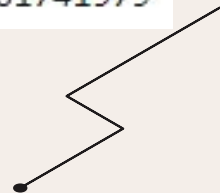


## ● LinearRegression

R2: 0.7068596784243815 // RMSE: 63740.94139671754 // MAE: 44609.07982362292

## ● RandomForest

R2: 0.8421088433733603 // RMSE: 18577.02011195424 // MAE: 11122.446061741979





# Axes d'améliorations



## Transformations

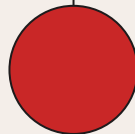
Enrichir la donnée

Trouver un moyen de gérer les outliers dans  
median\_house\_value



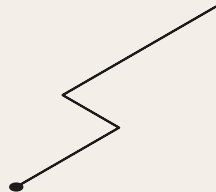
### RandomForest

Jouer sur les  
hyperparamètres pour  
améliorer le modèle



### LinearRegression

Tester d'autres modèles de  
régression linéaire



# Merci !

Des questions ?

