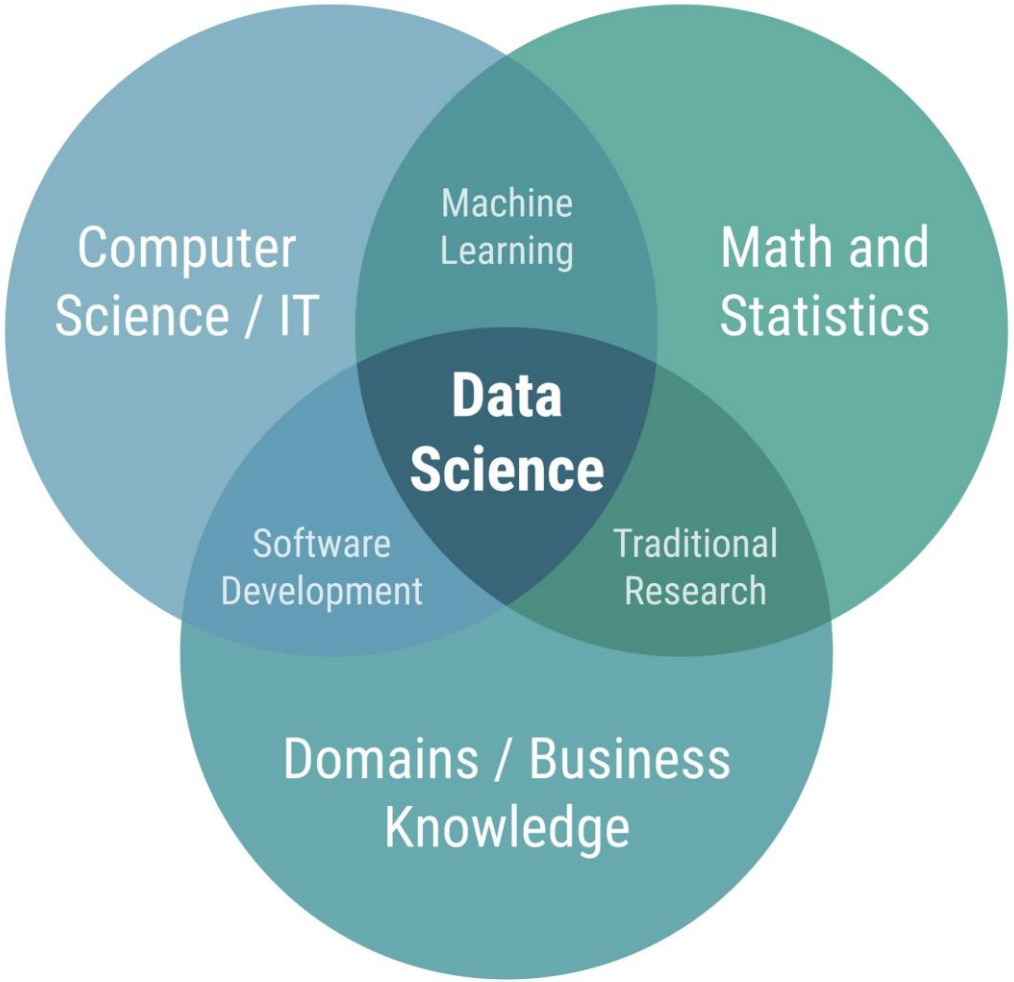
Data science

Sébastien AMALLA

Qu'est ce que la data science ?



Qu'est ce que la data science ?

- Science pluridisciplinaire regroupant :
  - Statistiques et mathématiques
  - Algorithmique
  - Expertise métier
- Objectif : extraire des informations à partir de données
- Les "données" sont les entrées bruts
- Les "informations" sont déduites du traitement des données
- Optimiser la prise de décisions, générer de la valeur

Trois domaines clés

- Mathématiques :
  - Statistiques
  - Probabilités
  - Modèles de machine learning
- Informatique (hacking skills) :
  - Python, R
  - Bases de données
  - Architecture et réseau
- Expertise métier

Objectif

- Prendre des décisions basées sur des données plutôt que sur l'intuition
- Prédire des comportements, des tendances
- Optimiser des processus
- Détecter des anomalies
- Santé : détection précoce de maladies, prédiction des épidémies
- Industrie : maintenance prédictive des machines, optimisation des chaines d'approvisionnement et des stocks
- Finance : prédictions boursières et analyses de risques

Idée de projet perso

- Téléchargez vos relevés de compte -> conversion en data
- Nettoyage et qualité des données
- Analyse descriptive des dépenses avec dashboard
- Détecter des abonnements oubliés
- Prédiction financières
- Décisions concernant l'épargne / l'investissement
- Optimisation du budget

Notion d'approche

- Plusieurs approches sont possibles selon les données et les objectifs
- Approche descriptive : Quoi ?
- Approche diagnostique : Pourquoi ?
- Approche prédictive : Que va-t-il se passer ?
- Approche prescriptive : Que devons-nous faire ?
- Approche expérimentale : Validation d'hypothèses

Approche descriptive (Quoi ?)

- Résumer et comprendre les données
- Analyse statistique :
  - Moyenne, médiane, écart-type, minimum, maximum, pourcentages, …
  - Corrélations entre variables
- Représentation :
  - Histogrammes, graphes, dashboards, …
- Identifier les tendances, motifs, anomalies, …

Approche diagnostique (Pourquoi ?)

- Comprendre pourquoi un phénomène ou un résultat se produit
- Exemple : les ventes ont chuté de 15% ce trimestre
- Analyse de causes et facteurs
- Comparaison par groupes : quels produits ont été impactés ?
  - -25% sur les produits électroniques, +8% sur les meubles, ...
- Regroupement par période : quelles semaines / jours ?
- Analyse de corrélation, heatmap
- Analyse de variance (ANOVA)

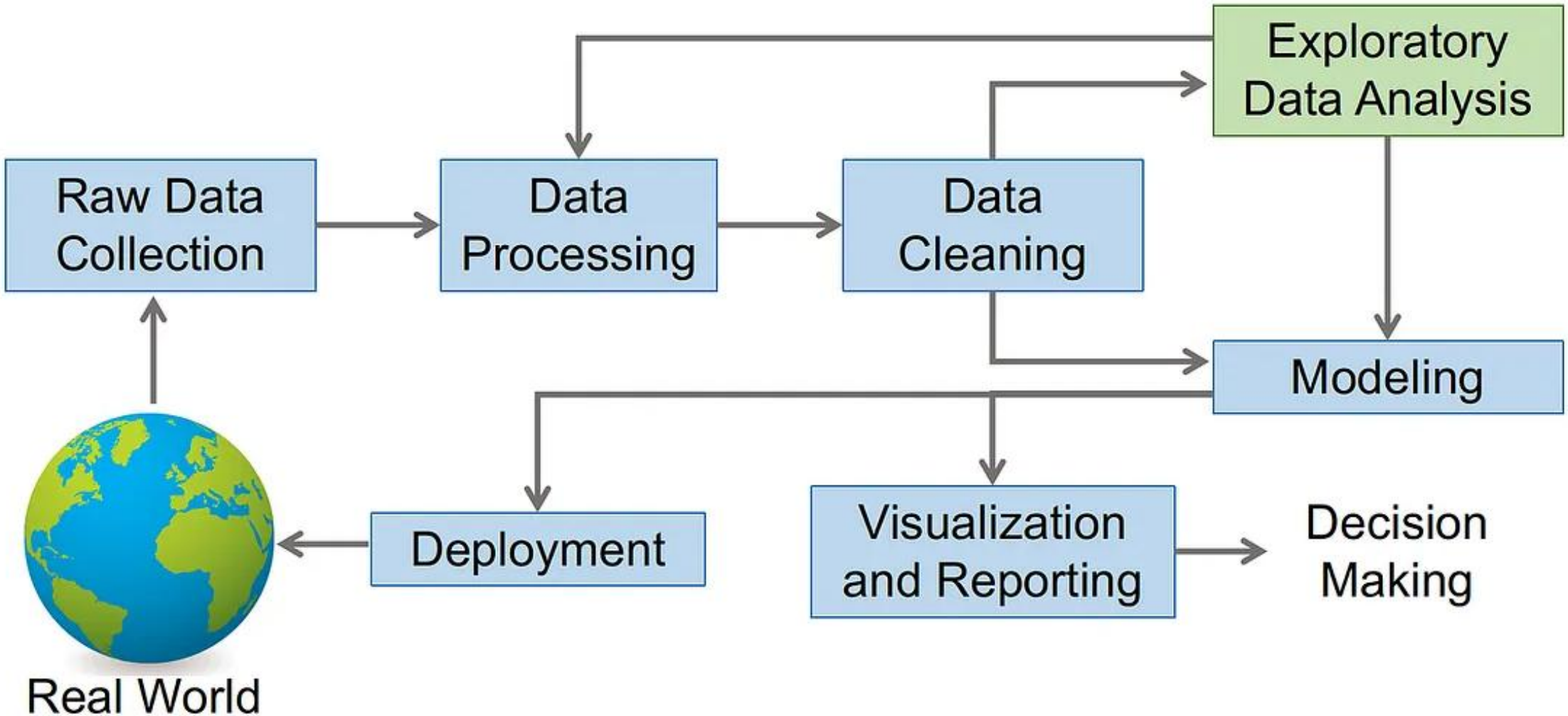
Approche prédictive (Futur ?)

- Statistiques et machine learning
- Apprentissage automatique:
  - Régressions : prédiction d'une valeur continue (prix, température)
  - Classifications : prédiction d'une catégorie (fraude ou pas)
  - Séries temporelles : tendances, saisonnalités, cyclicités, bruit, ...
- Choix du modèle
- Entraînement
- Validation
- Déploiement

Approche prescriptive : Que faire ?

- Définir des actions optimales pour un objectif donné
- Déterminer le meilleur stock à commander pour minimiser les coûts et les risques de rupture
- Déterminer un prix de vente
- Trajet d'une livraison
- Répartition des tâches dans une usine

Processus


Exploratory
Data Analysis
Raw Data
Collection
Data
Processing
Data
Cleaning
Modeling
Deployment
Visualization
and Reporting
Decision
Making
Real World

Datascience VS Machine Learning

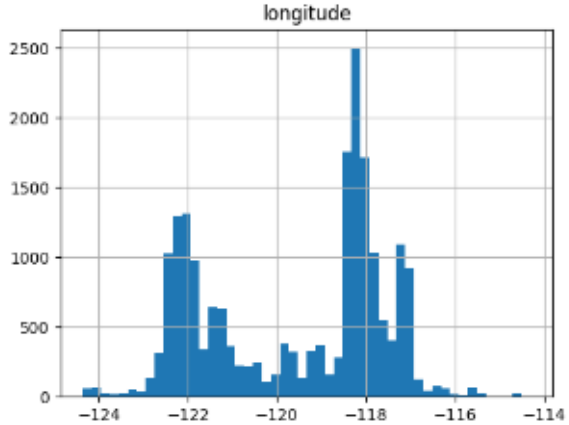
- Data Science = Discipline globale qui exploite les données pour créer de la valeur
- Machine learning = sous-domaine qui crée des modèles capable d'apprendre à partir des données
- Le ML est un outil que la datascience utilise, pas l'inverse
- La datascience concerne tout le cycle de vie des données

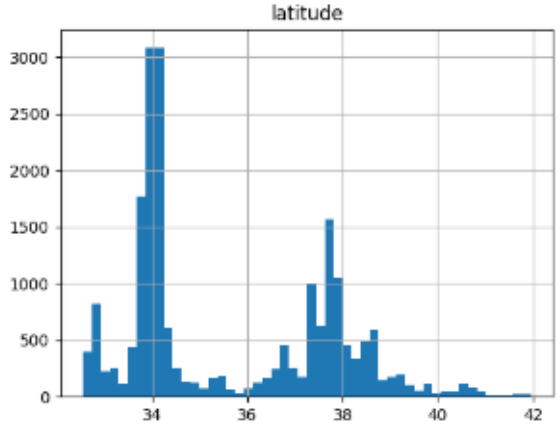
Projet concret

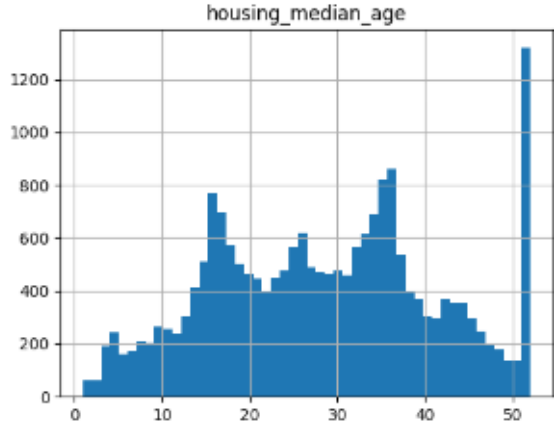
Prédiction du prix médian des maisons d'un quartier en Californie

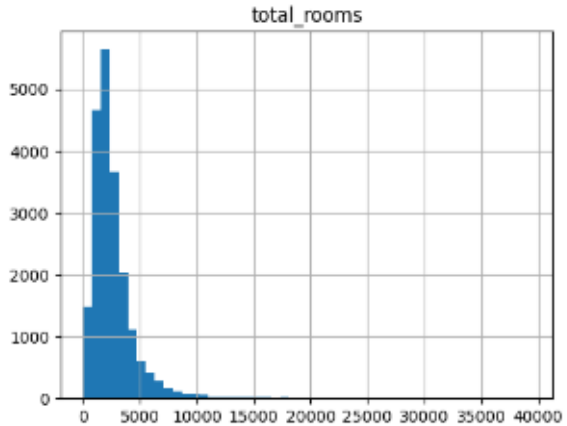
Marché de l'immobilier en Californie

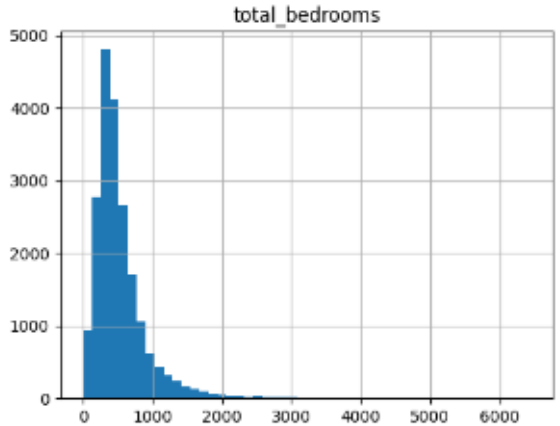
- Chaque ligne représente un disctrict
- On a la population totale
- Le nombre de maisons
- L'âge médian des maisons
- Le revenu médian des ménages
- La valeur médiane des maisons

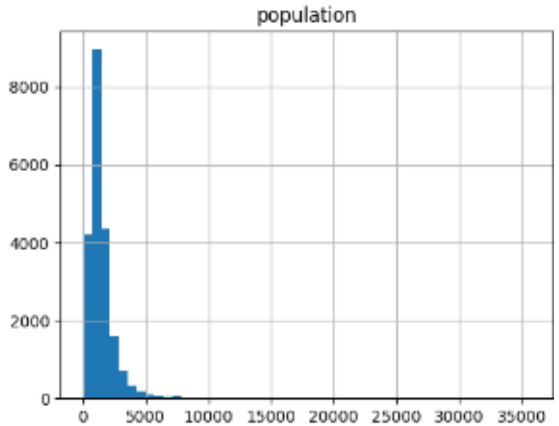

longitude
2500
2000
1500
1000
500
0
−124
−122
−120
−118
−116
−114

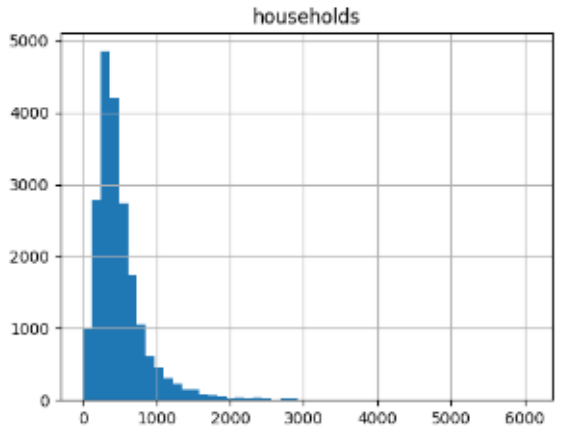

latitude
3000
2500
2000
1500
1000
500
0
34
36
38
40
42

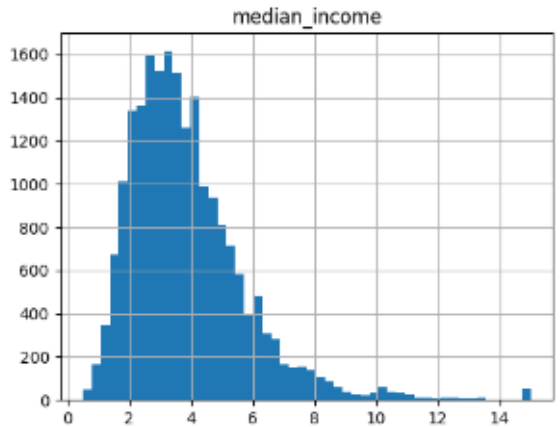

housing_median_age
1200
1000
800
600
400
200
0
0
10
20
30
40
50

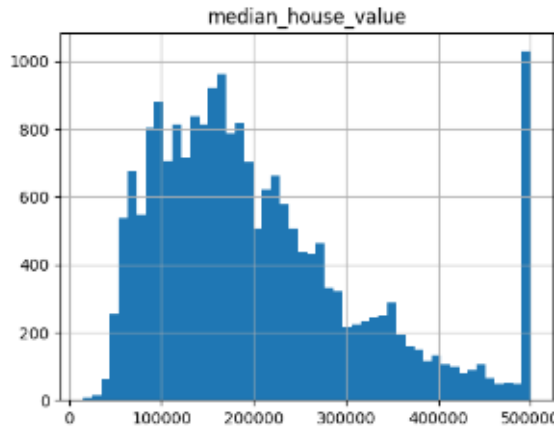

total_rooms
5000
4000
3000
2000
1000
0
0
5000
10000
15000
20000
25000
30000
35000
40000


total_bedrooms
5000
4000
3000
2000
1000
0
0
1000
2000
3000
4000
5000
6000


population
8000
6000
4000
2000
0
0
5000
10000
15000
20000
25000
30000
35000


households
5000
4000
3000
2000
1000
0
0
1000
2000
3000
4000
5000
6000


median_income
1600
1400
1200
1000
800
600
400
200
0
0
2
4
6
8
10
12
14

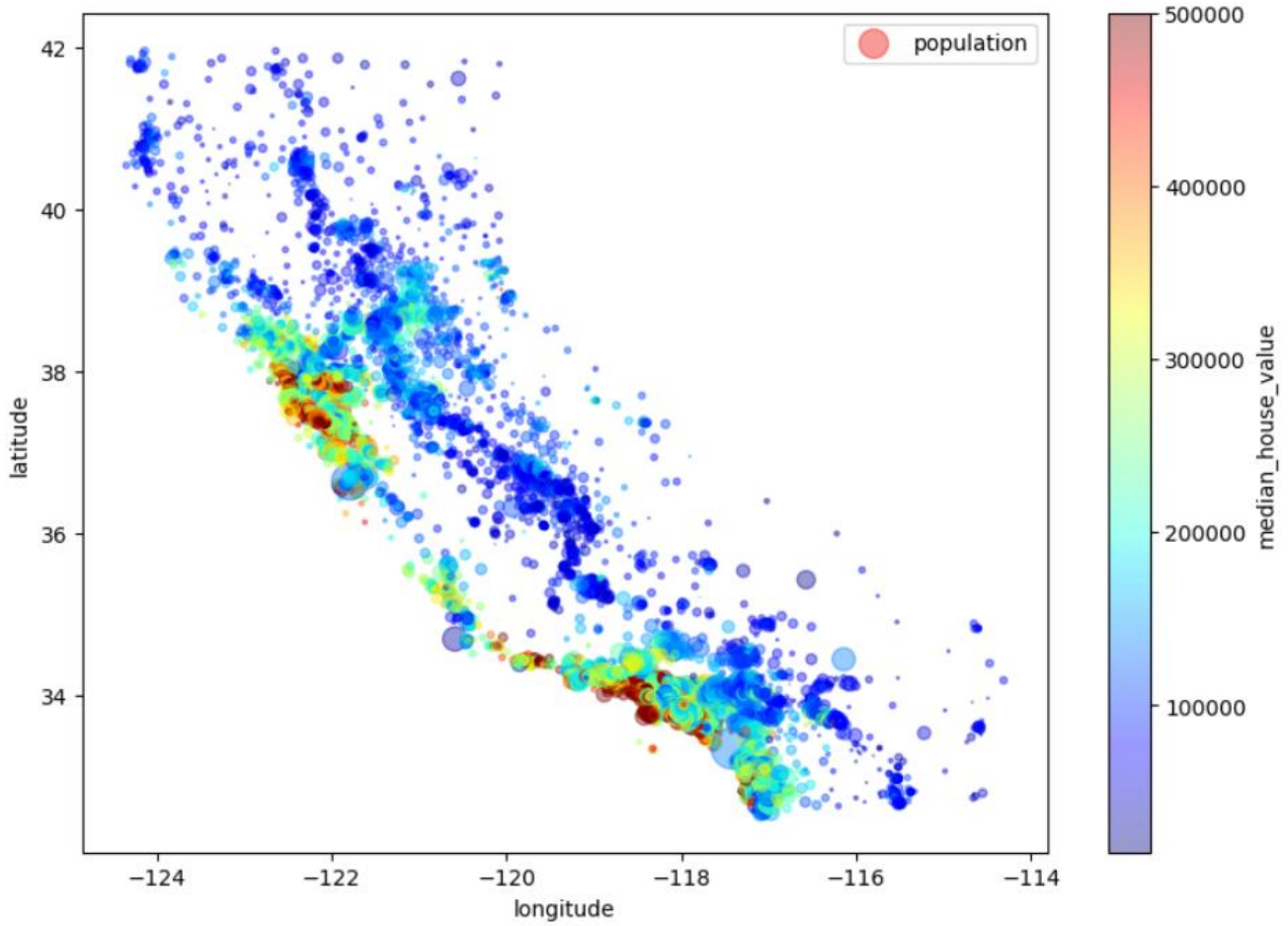

median_house_value
1000
800
600
400
200
0
0
100000
200000
300000
400000
500000

Que peut-on en déduire ?

- L'échelle de median_income n'est pas en dollar. Il faut demander à l'équiper qui a récolté les datas
- C'est une échelle qui va de 0,5 à 15 et chaque chiffre représente environ 10 000$
- housing_median_age et median_house_value ont également été bornées. Le ML va apprendre que les chiffres ne dépassent jamais ces limites
- Check avec les équipes métier : est-ce que les prédictions doivent dépasser les 500 000$?

Que peut-on en déduire ?

- Si oui, deux options :
  - Récolter des données plus précises
  - Enlever ces datas du set
- Les échelles des données sont très variables
- Le ML a besoin de données dans des échelles de même ordre de grandeur
- Les données s'étendent beaucoup plus à droite de leur médiane qu'à gauche
- "Normalisation des données" à faire plus tard



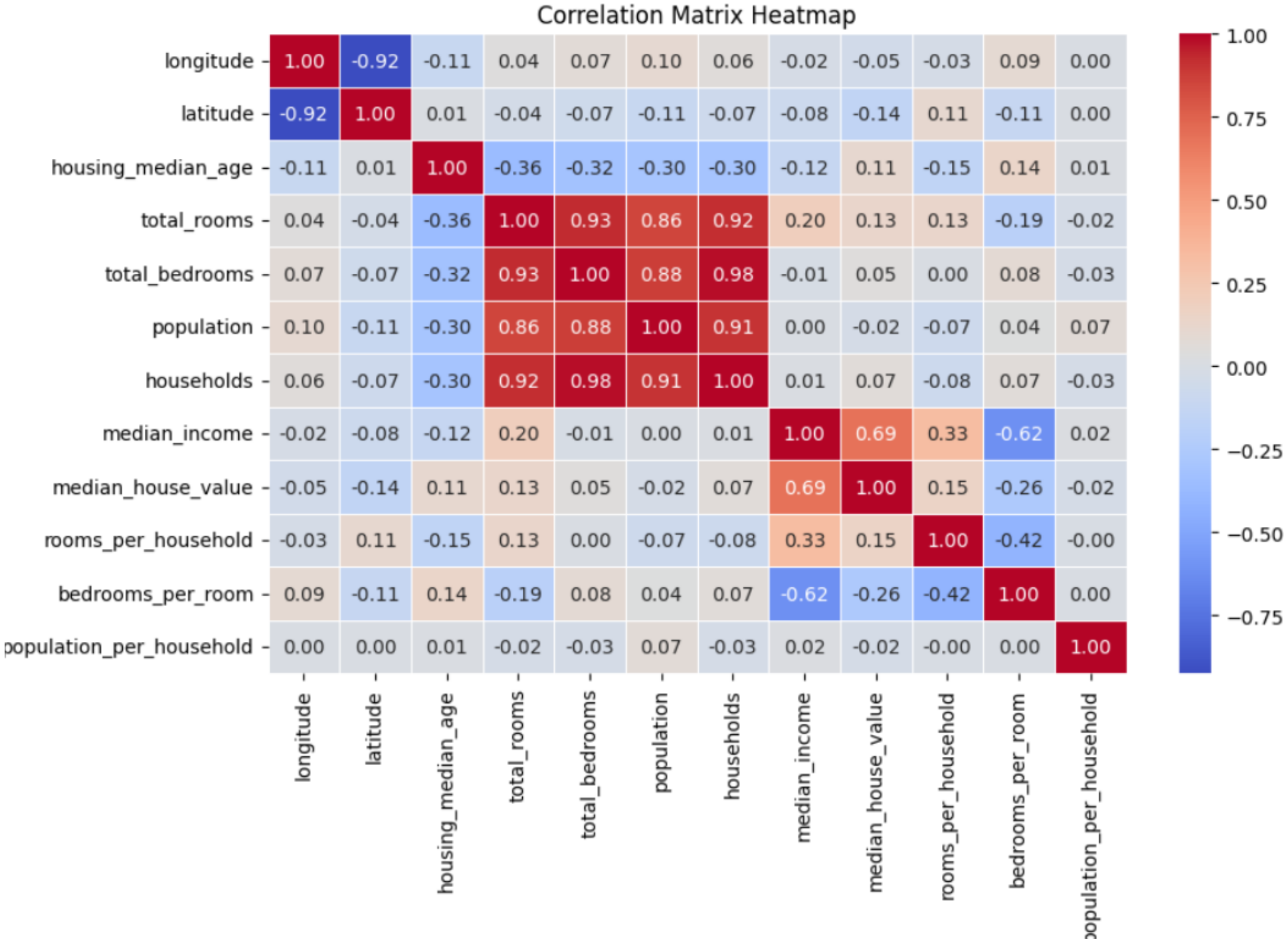
- Des zones sont plus représentées que d'autres
- Les prix sont plus élevés proche de l'océan

Corrélation linéaire entre variables

- Chaque variable peut être corrélée à toutes les autres
- Différence entre corrélation et causalité
- Différentes façon de la calculer : la plus commune est le coefficient de pearson
  - 0 = pas de corrélation
  - -1 = 100% de corrélation négative
  - +1 = 100% de corrélation positive
- Exemple "median_house_value" a une corrélation de -0,025 avec "population" et de +0,69 avec "median income"
- Que peut-on en déduire ?

Combinaisons d'attributs

- Nombre total de pièces et nombre total de foyer par district ne sont pas vraiment pertinents
- On va combiner les attributs pour en créer de nouveaux :
- Nombre de pièces par habitation
- Pourcentage de pièces qui sont des chambres
- Population par foyer
- On observe une grande corrélation entre bedrooms_per_rooms et median_house_value


Correlation Matrix Heatmap
longitude 1.00 -0.92 -0.11 0.04 0.07 0.10 0.06 -0.02 -0.05 -0.03 0.09 0.00
latitude -0.92 1.00 0.01 -0.04 -0.07 -0.11 -0.07 -0.08 -0.14 0.11 -0.11 0.00
housing_median_age -0.11 0.01 1.00 -0.36 -0.32 -0.30 -0.30 -0.12 0.11 -0.15 0.14 0.01
total_rooms 0.04 -0.04 -0.36 1.00 0.93 0.86 0.92 0.20 0.13 0.13 -0.19 -0.02
total_bedrooms 0.07 -0.07 -0.32 0.93 1.00 0.88 0.98 -0.01 0.05 0.00 0.08 -0.03
population 0.10 -0.11 -0.30 0.86 0.88 1.00 0.91 0.00 -0.02 -0.07 0.04 0.07
households 0.06 -0.07 -0.30 0.92 0.98 0.91 1.00 0.01 0.07 -0.08 0.07 -0.03
median_income -0.02 -0.08 -0.12 0.20 -0.01 0.00 0.01 1.00 0.69 0.33 -0.62 0.02
median_house_value -0.05 -0.14 0.11 0.13 0.05 -0.02 0.07 0.69 1.00 0.15 -0.26 -0.02
rooms_per_household -0.03 0.11 -0.15 0.13 0.00 -0.07 -0.08 0.33 0.15 1.00 -0.42 -0.00
bedrooms_per_room 0.09 -0.11 0.14 -0.19 0.08 0.04 0.07 -0.62 -0.26 -0.42 1.00 0.00
population_per_household 0.00 0.00 0.01 -0.02 -0.03 0.07 -0.03 0.02 -0.02 -0.00 0.00 1.00
longitude
latitude
housing_median_age
total_rooms
total_bedrooms
population
households
median_income
median_house_value
rooms_per_household
bedrooms_per_room
population_per_household
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

Création d'un set de test

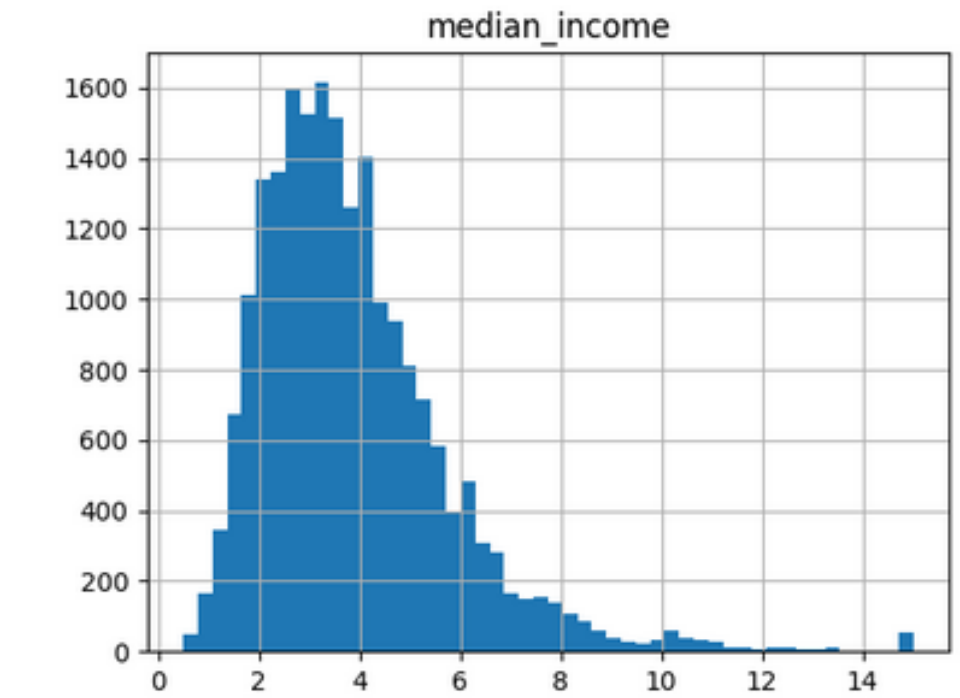
- Séparation des données en un set d'entraînement et un set de test
- En général 80% des données en train, 20% en test
- On veut pouvoir tester nos modèles avec des données sur lesquelles il ne s'est pas entraîné
- Façon simple : prendre 20% du dataset total au hasard

```
from sklearn.model_selection import train_test_split
```

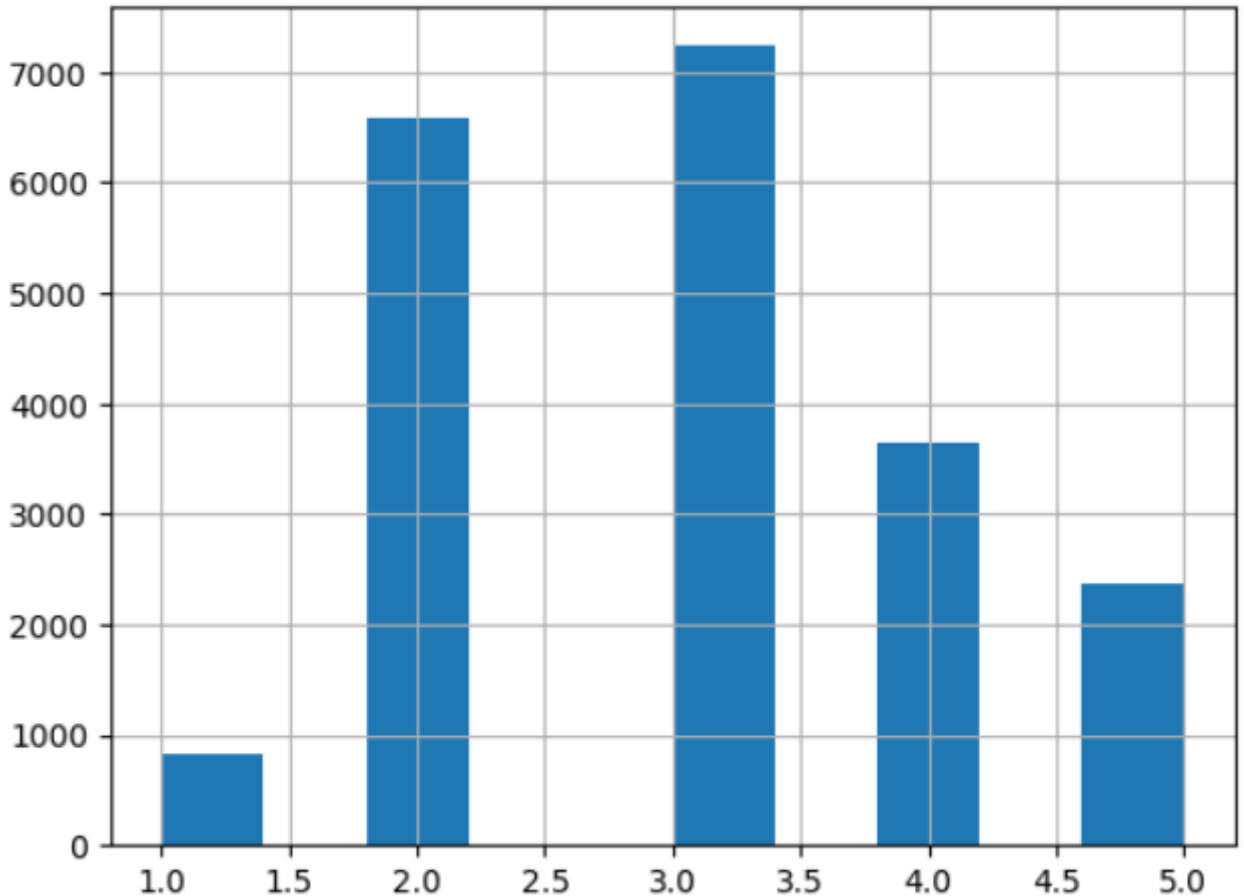
- Problème induit : il faut garder des échantillon représentatifs
- Pas un soucis pour un dataset large et homogène
- Étude sur 500 personnes
  - S'assurer d'avoir un échantillon représentatif des tranches d'âge en test et en train. Pas sûr si séparation faite au hasard

Stratified sampling

- Les données sont divisées dans des groupes homogènes : "strates"
- Identifier les strates par rapport au besoin métier
- L'équipe métier vous dit que le median_income est très important pour prédire le median_housing_prices
- Il faut s'assurer que les sets de train et de test soient représentatif de la distribution du set complet
- Il faut regarder de plus près la distribution de median_income



- Va de 0,5 à 15. La plupart se trouve entre 0,5 et 6, mais quelques unes vont bien plus loin
- On va le répartir en strates. Le but est d'avoir un nombre pertinent de strates : pas trop de strates, strates assez importantes
- On va faire 5 strates : la première de 0 à 1,5, la deuxième de 1,5 à 3, ..., la dernière de 6 à +infiny
- Fonction pandas "cut"


7000
6000
5000
4000
3000
2000
1000
0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0

Préparation des données pour le ML

- La variable à prédire est la "target feature"
- On va l'extraire de notre set d'entraînement, la mettre dans un tableau "labels"
- Entrée ayant "total_bedrooms" vide:
  - Soit on drop toutes les entrées vides
  - Soit on fill avec une valeur (zéro, moyenne, médiane, ...)
- On va choisit de fill par la médiane. Il va falloir calculer la médiane et la stocker ! Car il faudra fill le test_set avec la même valeur
- Objet "SimpleImputer" de sk-learn

sk-learn

- Utilisé dans dataiku
- Propose une multitude d'objets
- Ces objets sont à instancier et à "fit" sur les données
- Une fois "fit", ils peuvent "transform"
- Souvent, transforme sort un array numpy qui va devoir etre converti en DataFrame pandas
- Souvent, quand on doit fit sur des données puis les transformer, il existe une fonction "fit_transform" qui fait l'un puis l'autre

Variables catégorielles

- Les variables peuvent être non-numériques, comme du texte
  - Elles peuvent aussi être des images
- La variable "Ocean proximity" est un texte classant les entrée dans des catégories :
  - <1H OCEAN
  - INLAND
  - NEAR OCEAN
  - NEAR BAY
  - ISLAND
- Les modèles de machine learning apprennent sur des données numériques uniquement

Encoding

- Les variables catégorielles doivent être encodées
- Plusieurs choix différents, on en liste deux ici :
- **Ordinal encoding** = associe un chiffre à chaque catégorie :
  - 0 : <1H OCEAN
  - 1 : INLAND
  - 2 : NEAR OCEAN
  - 3 : NEAR BAY
  - 4 : ISLAND
- Induit une hiérarchie 4>3>2>1>0, des fois très problématique
- Très bien pour "good", "excellent", "bad" par exemple
- Pas d'appartenance à plusieurs catégories pour chaque entrée

Encoding

- **One-hot encoding :** On ajoute autant de colonnes que de catégories différentes
- Ces colonnes seront binaires : 0 ou 1
- Permet l'appartenance à deux catégories
- Pas de hiérarchie induite
- Peut grandement augmenter la taille du dataset si beaucoup de catégories
- Plusieurs autres encoding possibles…

Feature scaling

- Les modèles de ML ne fonctionnent pas bien si les variables ont des échelles très différentes
  - total_rooms : de 0 à 39 320
  - median_income : de 0 à 15
- Deux chois courants:
- Min-max scaling = Soustraire le min, diviser par max-min
  - On scale toutes les données dans une échelle allant de 0 à 1
- Standardization = Soustraire la moyenne et diviser par l'écart-type
  - Moyenne à zéro et écart-type de 1 sur une range infinie

Machine learning