Rapport

Mission 1:
Le dataset est complet, numérique et exploitable tel quel, mais présente de fortes hétérogénéités d'échelle et plusieurs valeurs extrêmes, notamment sur les variables liées à l'occupation et à la taille des logements. La variable cible est plafonnée, ce qui devra être pris en compte dans l'interprétation des résultats.

1. Structure générale du dataset

- **20 640 observations**, **9 variables**
- **Aucune valeur manquante**
- **Toutes les variables sont numériques (`float64`)**
- Mémoire faible (~1.4 MB) → dataset léger, rapide à manipuler

Très bon point de départ : pas de nettoyage lourd à prévoir.

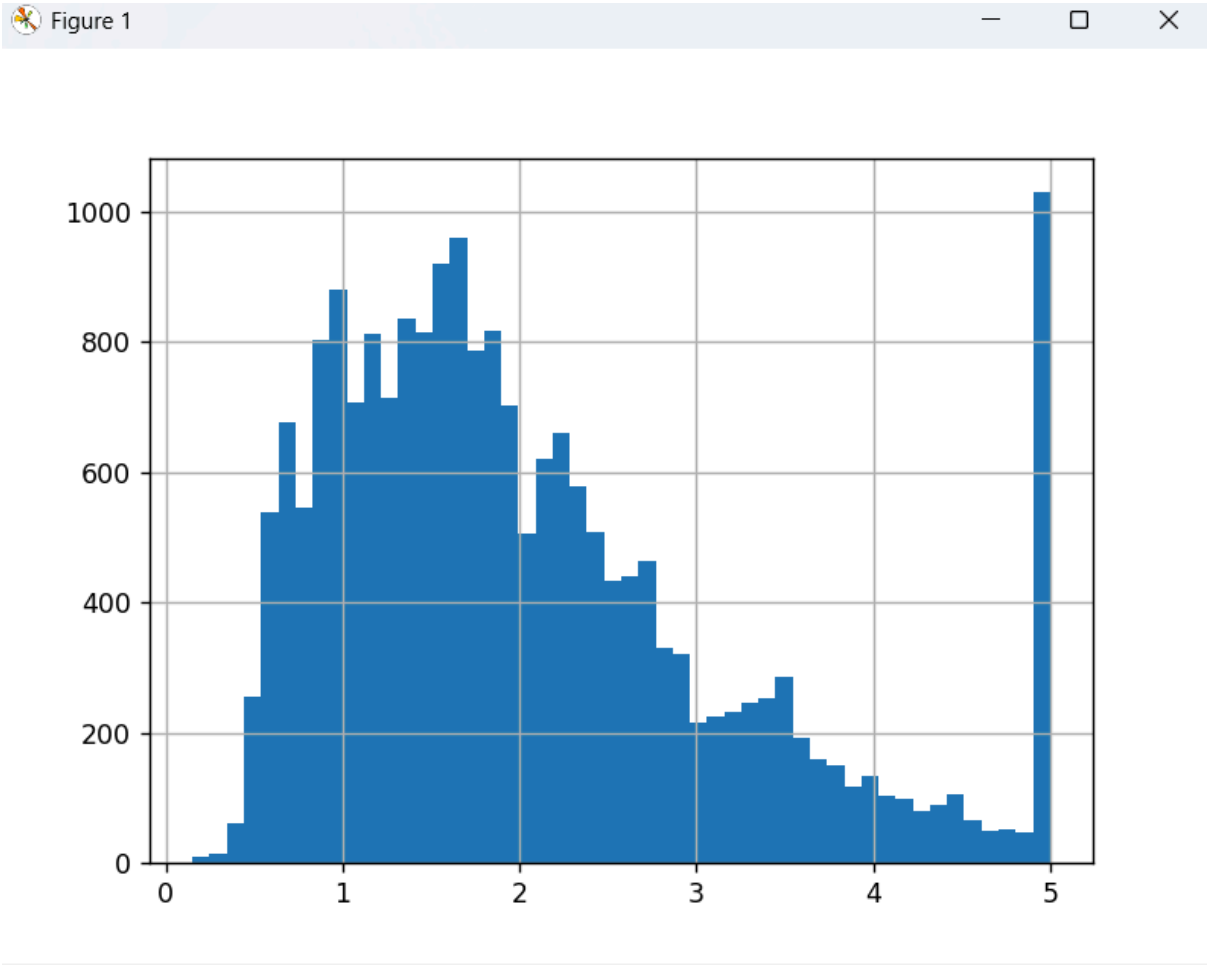
---

2. Variable cible : `MedHouseVal`

- Moyenne ≈ **2.07**
- Médiane ≈ **1.80**
- Max ≈ **5.00** (valeur plafond)

Distribution asymétrique à droite
Capping à 5 → important :

- biais possible en régression
- perte d'information pour les quartiers très chers Toutes les valeurs **au-dessus de 500k ont été tronquées à 5**



---

3. Variables explicatives – enseignements clés

`MedInc` (revenu médian)

- Large dispersion (0.5 → 15)
- Très probable **fort corrélat de `MedHouseVal`**
- Variable clé du modèle

Candidat n°1 pour expliquer le prix

`HouseAge`

- Entre **1 et 52 ans**
- Distribution assez équilibrée
- Pas de valeurs aberrantes

Effet probablement **non linéaire** (quartiers anciens ≠ toujours moins chers)

`AveRooms` & `AveBedrms`

- Moyennes raisonnables (~5.4 rooms, ~1.1 bedrooms)
- **Valeurs max énormes** (141 rooms, 34 bedrooms)

Présence de **forts outliers**

- quartiers atypiques
- erreurs d'agrégation possibles

`Population` & `AveOccup`

- Très forte dispersion
- `AveOccup` max ≈ **1243**

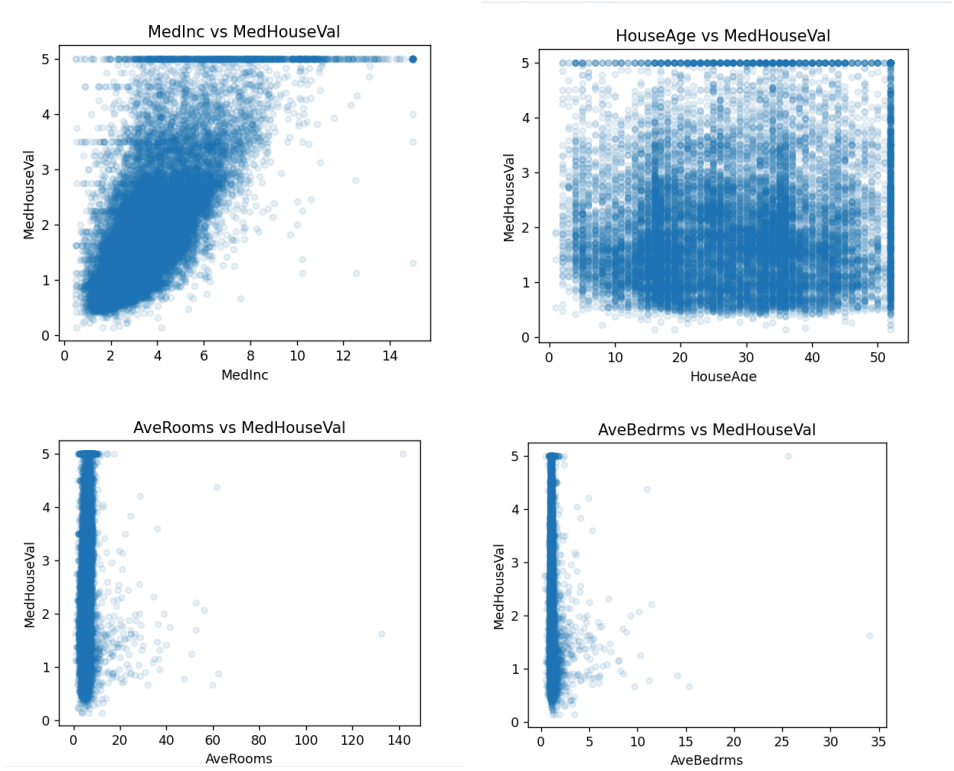
👉 Signal clair de **quartiers extrêmes**

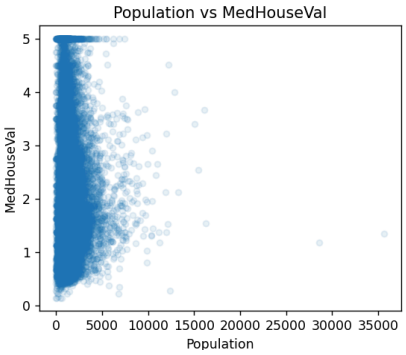
- zones très peu habitées
- colocation / logements collectifs
- risque de bruit pour le modèle

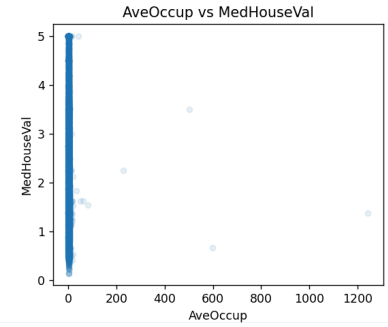
`Latitude` & `Longitude`

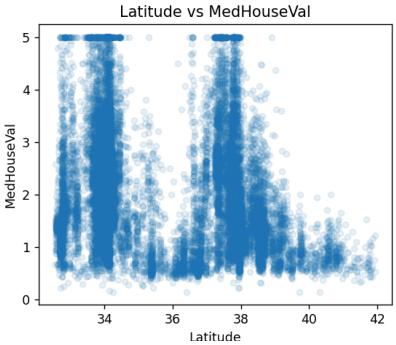
- Bornées sur la Californie
- Variables **géographiques très informatives**
- Relation non linéaire avec le prix

Comparaison variable:





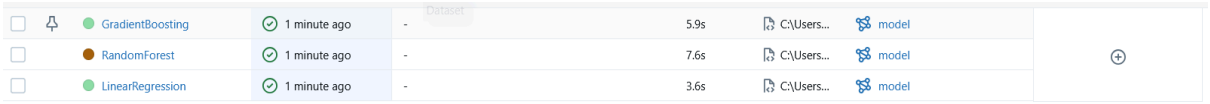




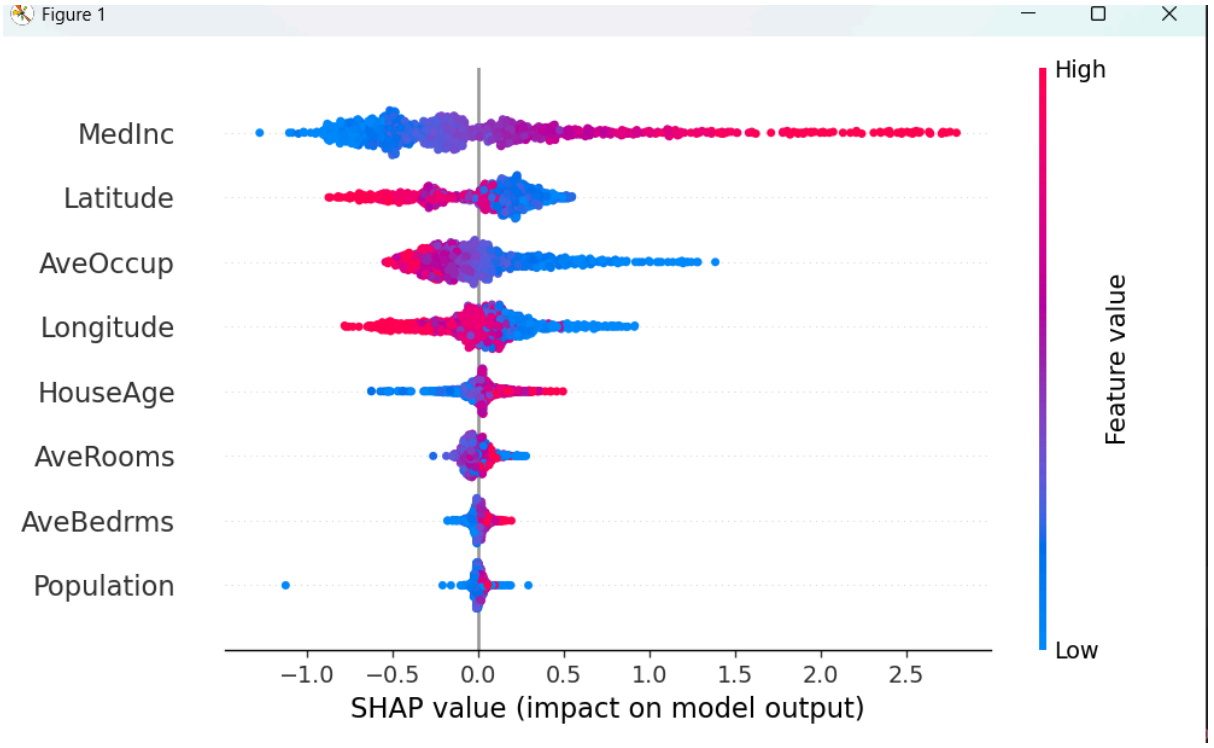
L’analyse exploratoire montre que le revenu médian (MedInc) et la localisation géographique (Latitude, Longitude) sont les variables les plus fortement associées à la valeur des logements. Les variables liées à l’occupation et à la taille des logements présentent de nombreux outliers et une relation faible ou bruitée avec la cible, nécessitant des transformations ou une régularisation. Enfin, la variable cible est plafonnée, ce qui introduit une saturation visible dans les relations et doit être prise en compte lors de la modélisation.

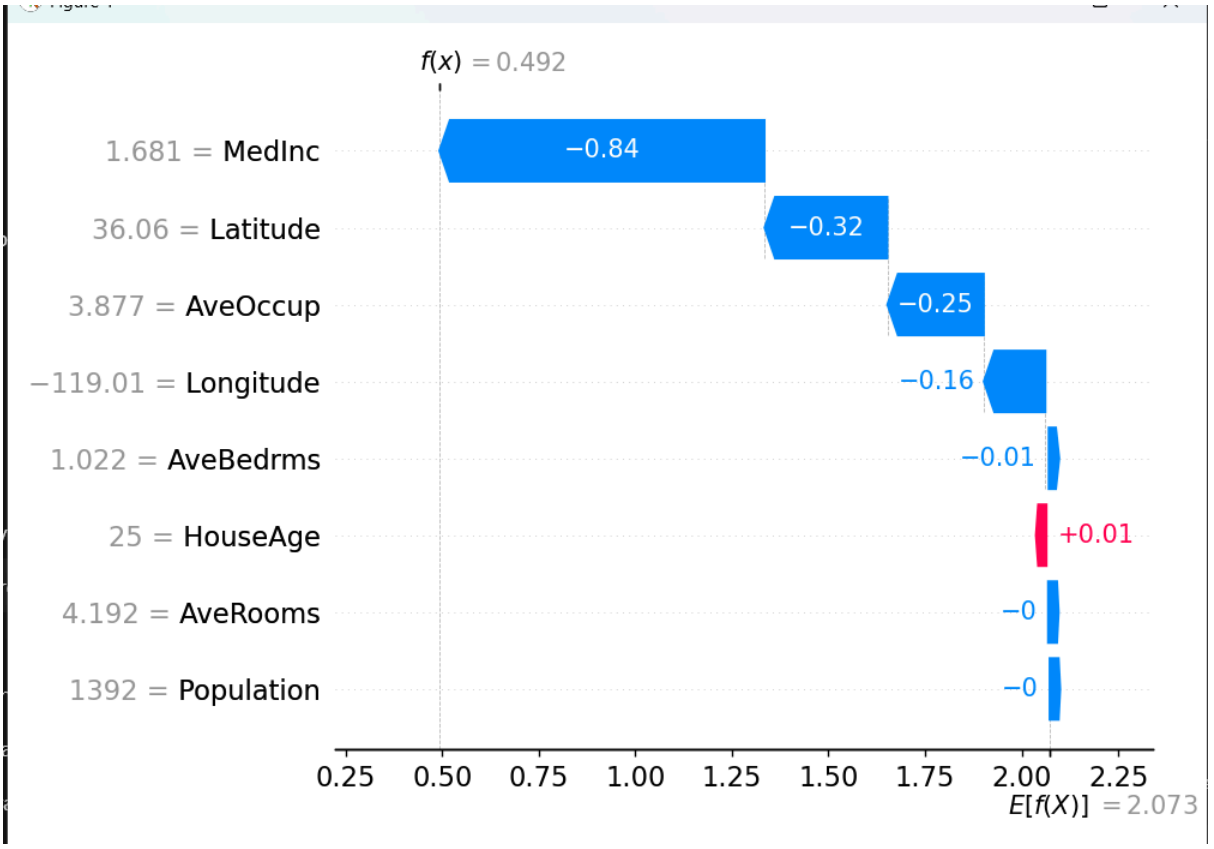
	MedInc	HouseAge	AveRooms	AveBedrms	Population	AveOccup	Latitude	Longitude	MedHouseVal
MedInc	1.000000	-0.119034	0.326895	-0.062040	0.004834	0.018766	-0.079809	-0.015176	0.688075
HouseAge	-0.119034	1.000000	-0.153277	-0.077747	-0.296244	0.013191	0.011173	-0.108197	0.105623
AveRooms	0.326895	-0.153277	1.000000	0.847621	-0.072213	-0.004852	0.106389	-0.027540	0.151948
AveBedrms	-0.062040	-0.077747	0.847621	1.000000	-0.066197	-0.006181	0.069721	0.013344	-0.046701
Population	0.004834	-0.296244	-0.072213	-0.066197	1.000000	0.069863	-0.108785	0.099773	-0.024650
AveOccup	0.018766	0.013191	-0.004852	-0.006181	0.069863	1.000000	0.002366	0.002476	-0.023737
Latitude	-0.079809	0.011173	0.106389	0.069721	-0.108785	0.002366	1.000000	-0.924664	-0.144160
Longitude	-0.015176	-0.108197	-0.027540	0.013344	0.099773	0.002476	-0.924664	1.000000	-0.045967
MedHouseVal	0.688075	0.105623	0.151948	-0.046701	-0.024650	-0.023737	-0.144160	-0.045967	1.000000

la matrice de corrélation conforte nos hypothèses avec probablement une relation non linéaire entre MedHouseVal et latitude longitude



Mission 3:





Les résultats montrent que la variable **MedInc** (revenu médian) est de loin la plus influente. Des valeurs élevées de revenu médian ont un impact fortement positif sur la prédiction du prix des maisons, tandis que des valeurs faibles contribuent à des prédictions plus basses. Cette relation est claire et cohérente sur l'ensemble du jeu de données et se vérifie également au niveau

individuel : dans l'exemple analysé, un revenu médian relativement faible entraîne à lui seul une baisse importante de la prédiction par rapport à la valeur moyenne du modèle.

Les variables géographiques, notamment **Latitude** et **Longitude**, jouent également un rôle important. Leur impact reflète des effets spatiaux marqués : pour l'observation considérée, la localisation géographique contribue négativement à la prédiction finale, confirmant que la position influence significativement la valeur des biens immobiliers en Californie. Ces effets restent toutefois plus modérés que celui du revenu médian.

La variable **AveOccup** (occupation moyenne) présente un impact négatif lorsque ses valeurs sont élevées, indiquant que des logements plus sur-occupés tendent à être associés à des prix plus faibles. Cet effet est visible à la fois globalement et localement, où une occupation supérieure à la moyenne réduit sensiblement la prédiction. D'autres variables comme **HouseAge**, **AveRooms** et **AveBedrms** ont un impact plus limité, souvent marginal ou non linéaire, ce qui souligne l'intérêt d'un modèle non linéaire pour capturer ces contributions fines.

Enfin, la variable **Population** apparaît comme peu influente dans les prédictions du modèle, avec des valeurs SHAP proches de zéro pour la majorité des observations, y compris dans l'exemple individuel analysé.

En conclusion, l'analyse SHAP confirme que le modèle repose principalement sur des facteurs socio-économiques et géographiques, en particulier le revenu médian, tandis que les variables démographiques globales jouent un rôle secondaire. L'analyse locale illustre en outre comment ces facteurs se combinent pour expliquer une prédiction individuelle, améliorant ainsi la compréhension et la transparence du modèle en complément des performances mesurées lors de la phase de modélisation.

ANNEXE

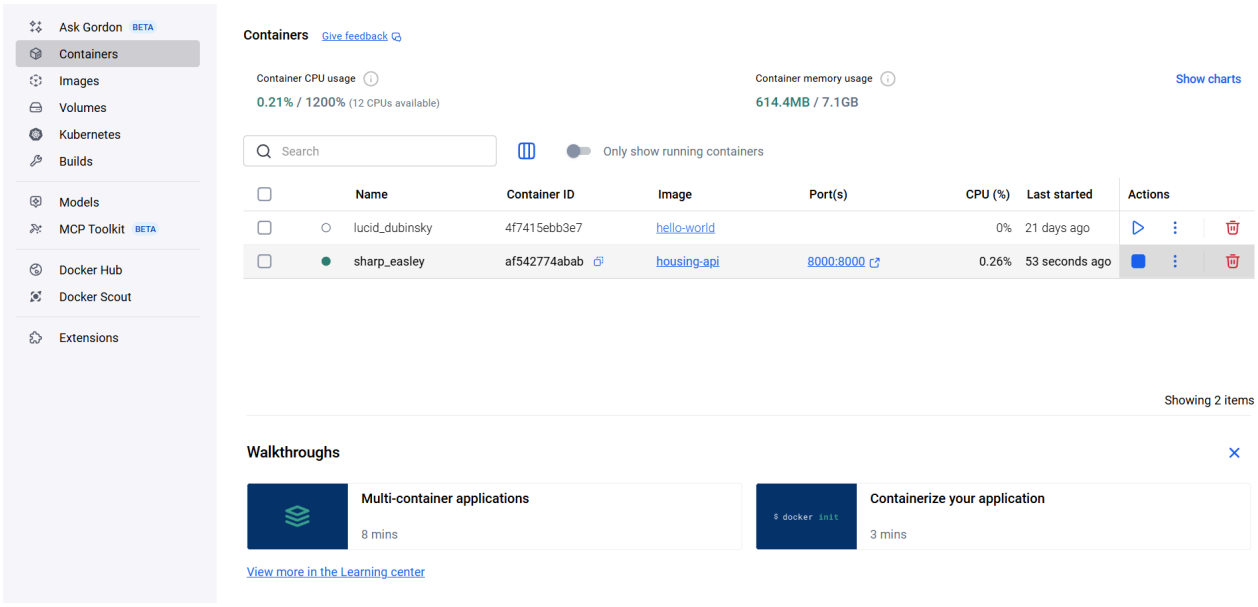
```
curl -X 'POST' \
  'http://127.0.0.1:8000/predict' \
  -H 'accept: application/json' \
  -H 'Content-Type: application/json' \
  -d '{
  "MedInc": 5.0,
  "HouseAge": 30.0,
  "AveRooms": 6.0,
  "AveBedrms": 2,
  "Population": 800.0,
  "AveOccup": 2,
  "Latitude": 14,
  "Longitude": -50
}'
```

Response body

```
{
  "prediction": 3.5133024333333274
}
```

Response headers

```
content-length: 33
content-type: application/json
date: Thu,25 Dec 2025 13:24:39 GMT
server: uvicorn
```



California Housing - API Client

MedInc

5.00 − +

HouseAge

30.00 − +

AveRooms

6.00 − +

AveBedrms

2 − +

Population

800.00 − +

AveOccup

2 − +

Latitude