Projet Modèle Linéaire

Antoine Legendre

November 2024

1 Introduction

L'objectif de ce projet est de prédire des prix immobiliers en Californie. Les données sont issues du **California Housing Dataset**. Les différentes variables de ce jeu de données sont regroupées par bloc de recensement :
- *MedInc* : revenu médian du bloc
- *HouseAge* : âge médian des logements du bloc
- *AveRooms* : nombre moyen de pièces par ménage
- *AveBedrms* : nombre moyen de chambres par ménage
- *Population* : population du bloc
- *AveOccup* : nombre moyen de personnes par ménage
- *Latitude* : latitude du bloc
- *Longitude* : longitude du bloc

La problématique de ce projet va donc être, en utilisant ces différentes variables, de prédire la valeur médiane des logements *MedHouseVal* d'un bloc.

2 Prise en main du jeu de données

2.1 Statistiques descriptives

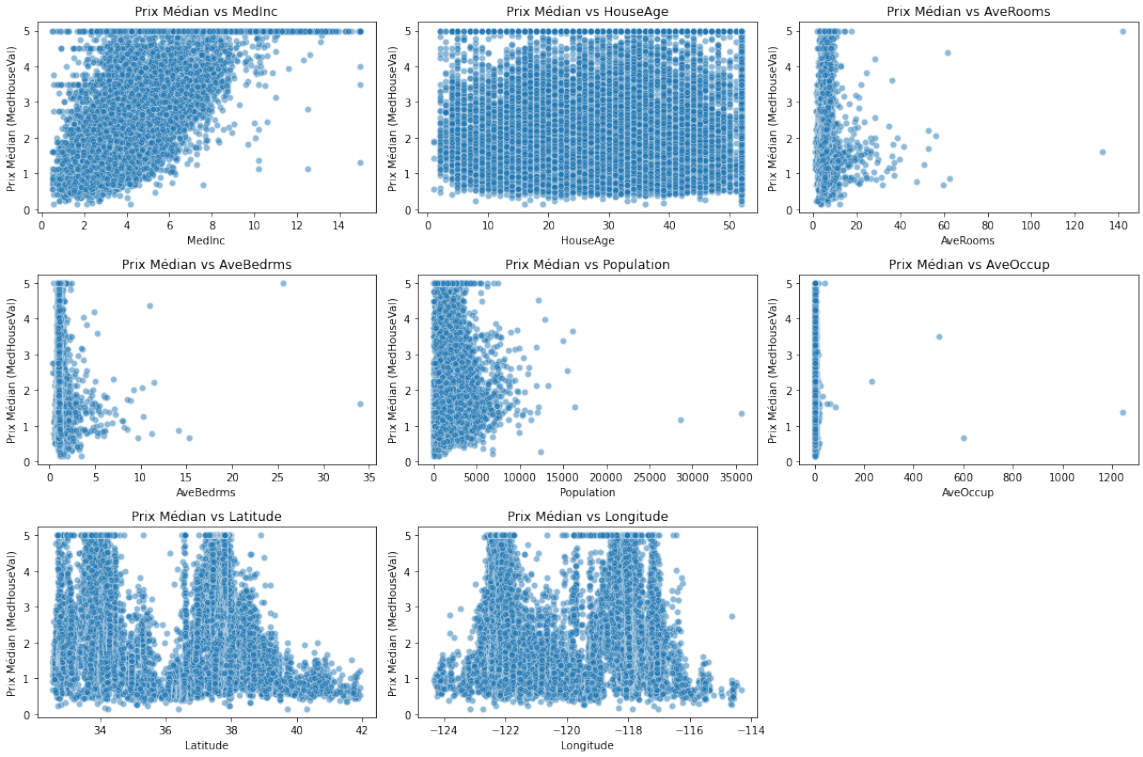
Tout d'abord, on va afficher les statistiques descriptives des différentes variables. Ces différentes statistiques comme les moyennes, mininimums, quartiles, maximums, etc. de chaque variable vont nous permettre d'avoir une vue d'ensemble sur les variables et de repérer des données potentiellement aberrantes si l'on remarque qu'elles sont trop éloignées des autres valeurs.

	MedInc	HouseAge	AveRooms	AveBedrms	Population	AveOccup	Latitude	Longitude	MedHouseVal
count	20640.0	20640.0	20640.0	20640.0	20640.0	20640.0	20640.0	20640.0	20640.0
mean	3.871	28.639	5.429	1.097	1425.477	3.071	35.632	-119.57	2.069
std	1.9	12.586	2.474	0.474	1132.462	10.386	2.136	2.004	1.154
min	0.5	1.0	0.846	0.333	3.0	0.692	32.54	-124.35	0.15
25%	2.563	18.0	4.441	1.006	787.0	2.43	33.93	-121.8	1.196
50%	3.535	29.0	5.229	1.049	1166.0	2.818	34.26	-118.49	1.797
75%	4.743	37.0	6.052	1.1	1725.0	3.282	37.71	-118.01	2.647
max	15.0	52.0	141.909	34.067	35682.0	1243.333	41.95	-114.31	5.0

On remarque que certaines variables (*AveRooms*, *AveBedrms*, *Population*, *AveOccup*) ont un maximum extrêmement élevé par rapport aux autres valeurs prises par ces variables. Pour *AveRooms* et *AveBedrms*, cela est sûrement lié à des blocs se situant dans des stations de vacances où il y a peu de ménages et de nombreux logements vides. Pour *Population* et *AveOccup*, on peut émettre l'hypothèse que cela est lié à un bloc où beaucoup de personnes résident dans le même logement comme un internat ou autre.

2.2 Prix médian des logements en fonction de chaque variable

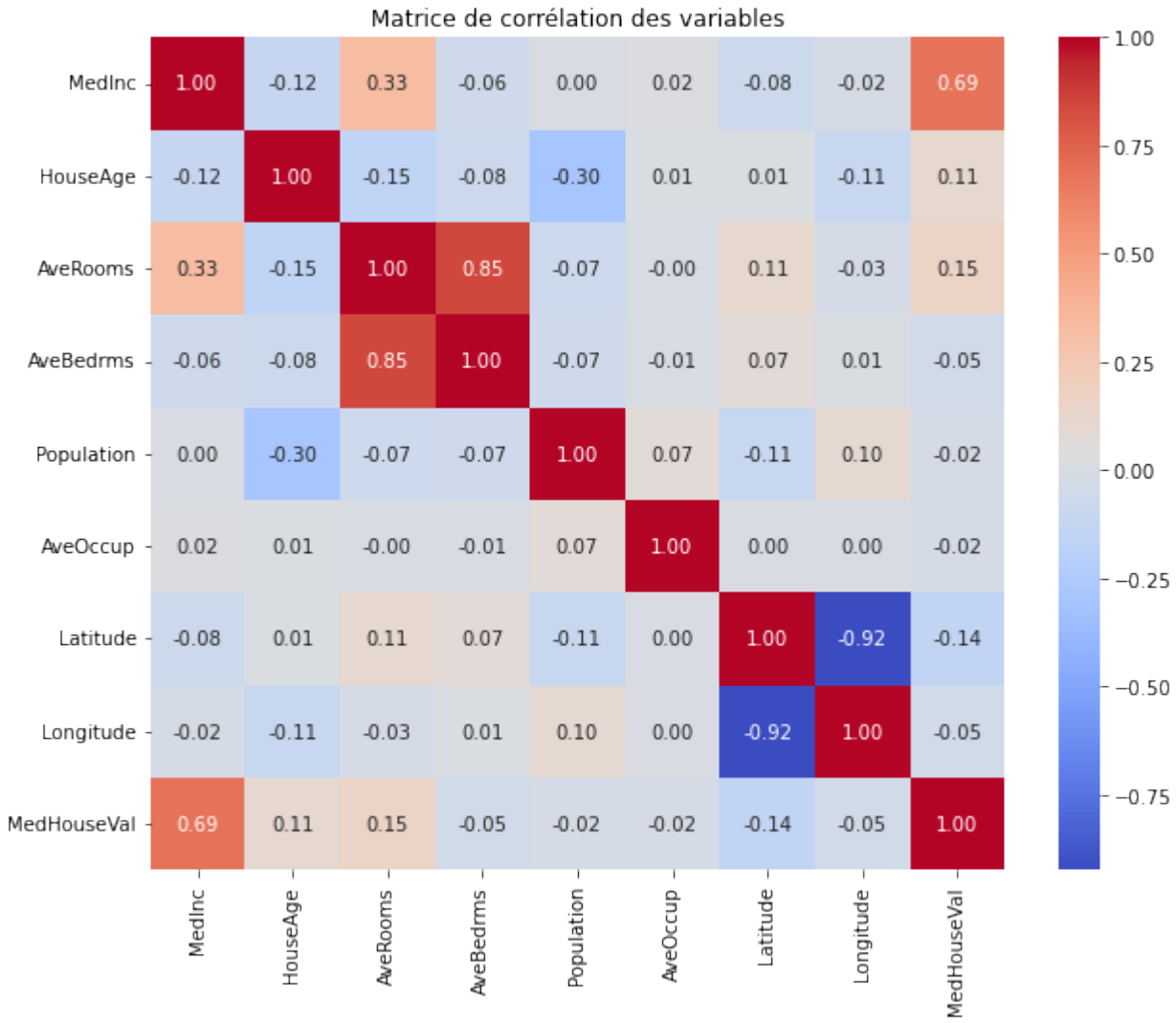
On va désormais réaliser les graphiques du prix médian des logements en fonction de chacune des variables. Cela va nous permettre d'observer la répartition des différentes variables et également de supposer des potentielles corrélations entre le prix médian des logements et les autres covariables.



On remarque à nouveau les valeurs extrêmes vues précédemment et l'on constate qu'il semblerait que les valeurs extrêmes de *AveRooms* coïncideraient avec celles de *AveBedrms* et celles de *Population* coïncideraient avec celles de *AveOccup* ce qui renforcent les hypothèses émises. On remarque également une forte corrélation entre *MedInc* et *MedHouseVal*. Enfin, pour certaines valeurs de *Latitude* et *Longitude* on remarque des valeurs plus élevées pour *MedHouseVal* même si cela ne devrait pas se traduire par une forte corrélation car *MedHouseVal* ne varie pas linéairement par rapport à *Latitude* ou *Longitude*.

2.3 Matrice de covariance

On va maintenant réaliser la matrice de covariance entre les différentes variables ce qui va nous permettre de voir quelles variables influent sur *MedHouseVal* et si certaines covariables sont fortement corrélées entre elles ce qui pourrait entraîner du sur-apprentissage et fausser l'estimation des moindres carrées.



La covariance entre *MedInc* et *MedHouseVal* est de 69% ce qui confirme ce que nous avions remarqué précédemment et ce qui signifie que 69% de la variance de *MedHouseVal* est expliquée par la variance de *MedInc*. La variable *MedInc*, représentant le revenu médian du bloc, explique donc grandement le prix médian des logements. Ensuite, on remarque, ce qui n'est peu surprenant, une forte corrélation (0.85) entre les covariables *AveRooms* et *AveBedrms* ainsi qu'une forte corrélation négative (-0.92) entre *Latitude* et *Longitude* ce qui s'explique par la forme de la Californie (diagonale du Nord-Ouest au Sud-Est). Il faudra donc prendre avec précaution les valeurs d'estimations des moindres carrés.

3 Modèles prédictifs

On va tout d'abord standardiser les covariables afin que les variables ayant de grandes valeurs numériques ne dominent pas l'ajustement du modèle. Tous les coefficients donnés dans cette partie 3 seront donc les coefficients pour des données standardisées.

3.1 Premier modèle prédictif à l'aide du critère AIC

On va créer notre premier modèle prédictif en partant de toutes les covariables telles quelles sans introduire de pénalisation ou autres covariables pour observer les potentielles qualitées prédictives de ce modèle. On cherche un bon modèle prédictif, on va donc utiliser le critère AIC afin de sélectionner les covariables expliquant le mieux le prix médian des maisons.

On obtient alors un R^2 de 0,606, ce qui signifie que le modèle explique $60,6\%$ de la variance de *MedHouseVal*. Ce modèle est donc modérément performant pour estimer les valeurs médianes des prix des maisons.
Le modèle a sélectionné toutes les covariables hormis la variable *Population*. Hormis la population du bloc, les variables expliquent donc toutes plus ou moins significativement le prix médian des logements d'un bloc. Le modèle a fourni les coefficients suivants pour les covariables sélectionnées :

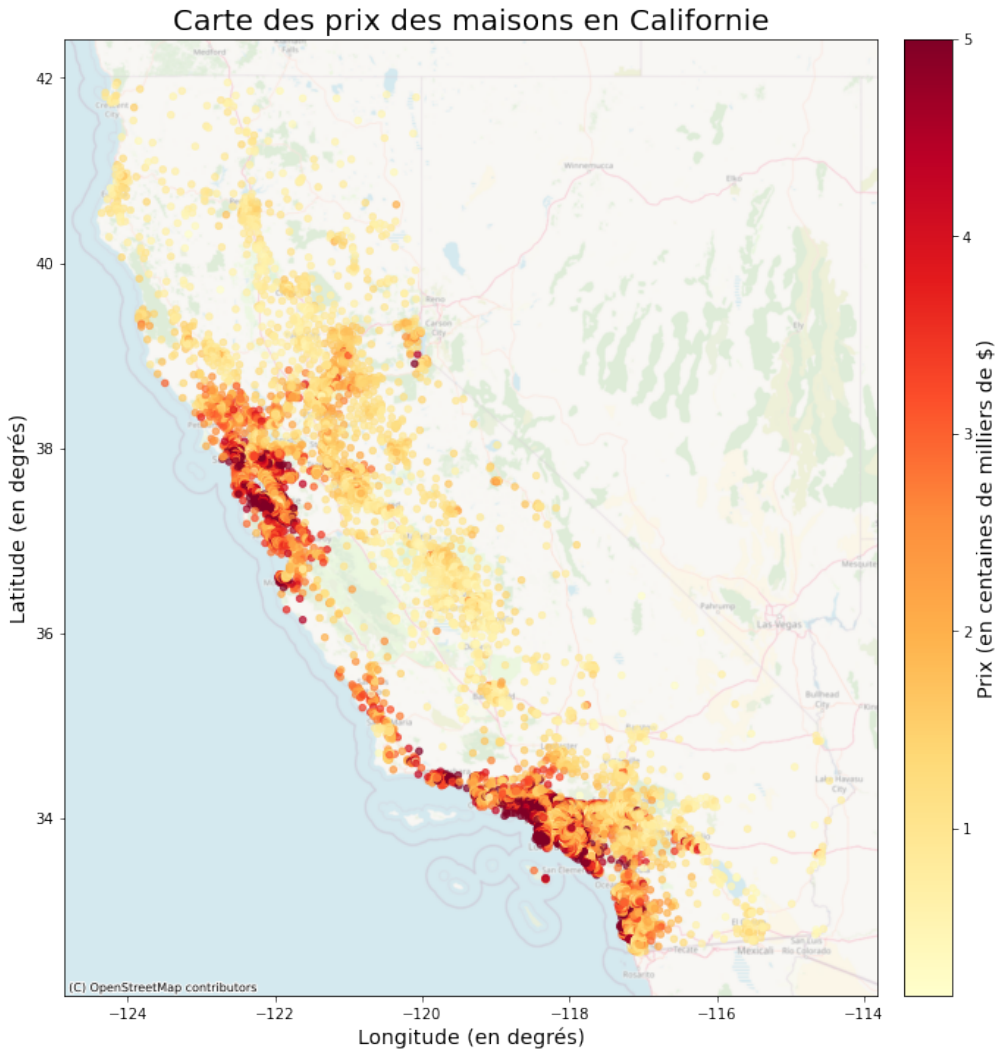
Covariable	**Coefficient**
Intercept	2,069
MedInc	0,830
HouseAge	0,120
Latitude	-0,899
Longitude	-0,870
AveBedrms	0,306
AveRooms	-0,265
AveOccup	-0,040

TABLE 1 – Coefficients de la régression linéaire.

On remarque que la variable ajoutée en premier et ayant un des coefficients les plus élevés en valeur absolue est *MedInc* comme on pouvait s'y attendre. En calculant le R^2 ajusté par validation croisée, on obtient un R^2 de 0.554 ce qui est légèrement inférieur au R^2 trouvé précédemment (0.606) et ce qui s'explique par un sur-apprentissage sûrement dû à la forte corrélation entre *Latitude* et *Longitude*.

3.2 Nouvelles variables

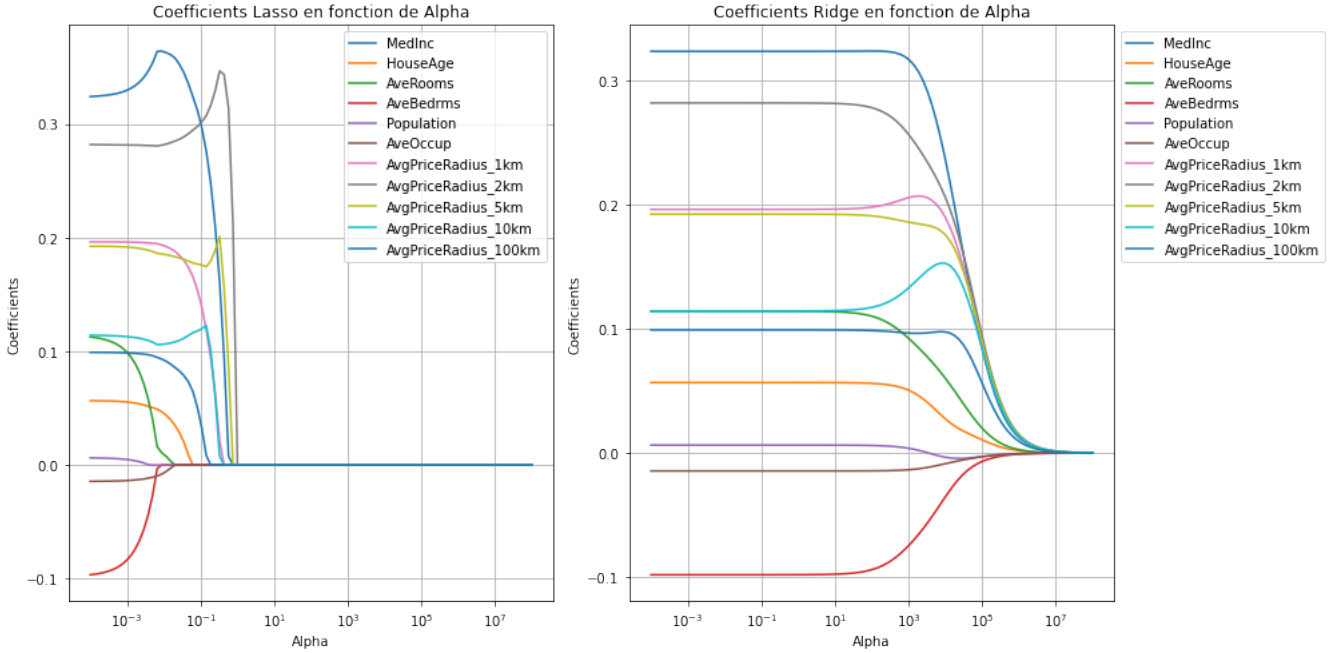
On a vu que *Latitude* et *Longitude* influaient sur *MedHouseVal* et étaient fortement corrélées entre elles, il faudrait donc trouver un moyen de créer une nouvelle variable regroupant ces deux variables et qui expliquerait mieux le prix médian des maisons. Pour déterminer quelle nouvelle variable serait la plus adaptée pour remplacer ces deux variables nous allons réaliser une carte de la Californie qui va permettre de visualiser la répartition géographique des différents blocs et des prix médians des maisons à l'aide d'un code couleur.



On remarque que les points les plus foncés, c'est à dire les blocs où le prix médian des maisons est le plus élevé sont majoritairement regroupés aux mêmes endroits. On va donc pouvoir créer de nouvelles variables contenant les prix médians des logements situés dans les blocs à moins de x km de ce bloc (ce bloc exclu). Après quelques tests, on se rend compte qu'on obtient la variable expliquant le mieux *MedHouseVal* en prenant une distance x aux alentours de 3 km. Cependant, on se rend compte qu'en créant plus de covariables, c'est à dire en prenant plusieurs distances x le modèle explique encore mieux *MedHouseVal*. On choisit donc de prendre $x = [1, 2, 5, 10, 100]$ ce qui va donc créer les nouvelles variables *AvgPriceRadius_ 1km*, *AvgPriceRadius_ 2km*, etc. L'ajout de nouvelles variables va rendre le modèle plus précis mais il sera alors plus complexe et moins efficace.

3.3 Régressions Ridge et Lasso

On va maintenant réaliser une régression Ridge pour obtenir un modèle sans sur-apprentissage ainsi qu'une régression Lasso pour obtenir un modèle sans sur-apprentissage avec moins de covariables et donc moins complexe. Tout d'abord on va réaliser des graphiques montrant l'évolution des valeurs des coefficients des covariables en fonction du coefficient de pénalisation alpha afin de déterminer quelle valeur de alpha choisir pour chaque régression.



Au vu de ces graphiques, on choisit $alpha = 10^{-1}$ pour la régression Lasso afin d'annuler certaines covariables tout en gardant un modèle précis et $alpha = 10^{2}$ pour la régression Ridge ce qui va permettre de garder la précision du modèle. On remarque également sur le graphique de la régression Ridge que les $AvgPriceRadius\_xkm$ se compensent entre eux lorsque l'un d'entre eux s'annulent ce qui démontre une légère corrélation entre ces variables ce qui n'a rien d'étonnant.

En prenant les valeurs de alpha citées précédemment, on obtient les coefficients suivants :

Covariable	**Lasso**	**Ridge**
MedInc	0.301	0.324
HouseAge	0.000	0.056
AveRooms	0.000	0.110
AveBedrms	-0.000	-0.094
Population	-0.000	0.006
AveOccup	-0.000	-0.014
AvgPriceRadius_ 1km	0.143	0.198
AvgPriceRadius_ 2km	0.301	0.278
AvgPriceRadius_ 5km	0.176	0.191
AvgPriceRadius_ 10km	0.119	0.117
AvgPriceRadius_ 100km	0.038	0.099
Intercept	2.069	2.068

TABLE 2 – Coefficients des modèles Lasso et Ridge.

La régression Lasso nous a permis de passer d'un modèle à 11 covariables à 6 covariables. De plus on obtient un R^2 de 0.800 et un R^2 ajusté par validation croisée de 0.796 ce qui montre de bonnes qualités prédictives et presque aucun sur-apprentissage. La régression Lasso nous fournit donc un modèle fiable et assez simple (seulement 6 covariables). On remarque que les seules covariables restantes avec la régression Lasso sont la variable *MedInc*, ainsi que les variables *AvgPriceRadius_ xkm*, toutes les autres variables du jeu de données ont des coefficients nuls. Le prix médian des maisons au sein d'un bloc s'explique donc majoritairement par le revenu médian du bloc et par les prix médians des logements situés dans les blocs proches de ce bloc. Enfin, avec la régression Ridge on remarque également que ces mêmes variables ont une grande influence et on obtient un R^2 de 0.812 et un R^2 ajusté par validation croisée de 0.812 également ce qui montre de bonnes qualités prédictives et aucun sur-apprentissage.

3.4 Nouveau modèle prédictif à l'aide du critère AIC

On va maintenant chercher un nouveau modèle prédictif, à l'aide du critère AIC, avec nos nouvelles variables et on obtients lors ces résultats :

Covariable	**Coefficient**
Intercept	2.069
AvgPriceRadius_ 2km	0.282
MedInc	0.319
AvgPriceRadius_ 10km	0.123
AvgPriceRadius_ 1km	0.195
AvgPriceRadius_ 100km	0.092
AvgPriceRadius_ 5km	0.192
HouseAge	0.056
AveOccup	-0.015
AveRooms	0.112
AveBedrms	-0.096
Population	0.008

TABLE 3 – Coefficients de la régression linéaire.

On remarque que la régression linéaire à l'aide du critère AIC a sélectionné toutes les variables. Les valeurs des coefficients de cette régression sont assez proches des valeurs des coefficients de la régression Ridge. On obtient également, comme pour la régression Ridge, un R^2 de 0.812 mais le R^2 ajusté par validation croisée est ici de 0.797 sûrement à cause d'un léger sur-apprentissage dû à la corrélation entre les différentes variables *AvgPriceRadius_ xkm*. La régression Ridge est donc la plus performante de ces régressions pour prédire *MedHouseVal*.

3.5 Implémentation des modèles

On a vu que le modèle fourni par la régression Ridge était le plus performant pour prédire le prix médian des maisons, on va donc implémenter ce modèle afin qu'il puisse être utilisé. Pour cela, il va tout d'abord falloir "déstandardiser" les coefficients, c'est-à-dire les ajuster à des données non standardisées. On demande ensuite à l'utilisateur de rentrer des valeurs pour chacunes des variables et le code va renvoyer une prédiction du prix médian des maisons dans le bloc. En rentrant les informations du premier bloc du **California Housing Dataset** on obtient alors un prix de 4.608 (en centaines de milliers de dollars) ce qui est assez proche du prix réel de 4.526.
On a également vu que la régression Lasso fournissait un modèle plus simple avec de bonnes qualités prédictives, on va donc l'implémenter également. En reprenant l'exemple du bloc utilisé précédemment et en rentrant seulement les 6 covariables nécessaires pour ce modèle, on obtient alors un prix de 4.272 ce qui reste raisonnablement proche du prix réel (4.526).

4 Conclusion

En conclusion, nous avons réussi à produire un modèle à 11 covariables avec un R^2 de 0.812 et un R^2 ajusté par validation croisée de 0.812 également et un modèle à 6 covariables avec un R^2 de 0.800 et un R^2 ajusté par validation croisée de 0.796. Ces deux modèles ont donc de bonnes qualités prédictives et un sur-apprentissage faible voir nul et permettent donc d'avoir des réponses fiables à la problématique posée qui était de prédire la valeur médiane des logements d'un bloc.
Cependant, un R^2 de 0.812 laisse place à des améliorations et peut laisser penser que d'autres facteurs influent sur le prix médian des logements. Par exemple, on remarque sur la carte de la Californie que les points correspondants aux prix médians les plus élevés se situent en fait au niveau des villes de San Francisco et Los Angeles, on pourrait donc penser à récupérer les coordonées de ces villes et créer de nouvelles variables sur la distance entre ces villes et le bloc.