

# Rapport

## Etude sur les prix de logements

### Techniques de régression avancées

---



Réalisé par :

LEPERCQ Louise

NITA Carmen Andreea

SARR Adja

THIOUNE Magatte Niang

Encadré par :

DELSART Virginie

MORGE Maxime

---

## Sommaire

<b>Introduction</b>	<b>4</b>
<b>1. Présentation de l'équipe</b>	<b>5</b>
<b>2. Mise en place de différent outils et organisation</b>	<b>6</b>
2.1. Mise en place du Trello	6
2.2. Mise en place du Git	7
2.3. Organisation	9
<b>3. Présentation des données</b>	<b>11</b>
3.1. Description des données	11
3.2. Hypothèses	11
<b>4. Mise en qualité des données</b>	<b>12</b>
4.1. Valeurs manquantes	12
4.2. Valeurs aberrantes	14
<b>5. Feature engineering</b>	<b>16</b>
5.1. Ajout de différentes sources externes	16
<b>5.2. Préparation de la base de données</b>	<b>19</b>
5.2.1. Modifications de variables	19
5.2.2. Création de variables	20
<b>6. MCD (Modèle Conceptuel de Données)</b>	<b>23</b>
<b>7. Statistiques descriptives</b>	<b>25</b>
7.1. Statistiques univariées	26
7.2. Statistiques bivariées	30
<b>8. Pertinence des variables</b>	<b>36</b>
<b>9. Relation entre les variables</b>	<b>36</b>
9.1. Coefficient de corrélation	36
9.2. Procédure ANOVA	37
<b>10. Préparation des données pour le Machine Learning</b>	<b>40</b>
<b>11. Problème rencontré pour le Machine Learning</b>	<b>41</b>
<b>12. Construction et évaluation des modèles</b>	<b>42</b>
12.1. Modèle 1 - Régression linéaire	42
12.2. Modèle 2 - Lasso	42
12.3. Modèle 3 - Ridge	43
12.4. Modèle 4 - Kernel Ridge	43
12.5. Modèle 5 - ElasticNet	44
12.6. Modèle 6 - Arbre de décision régressif	44
12.7. Modèle 7 - Machine à Vecteurs de Support (SVM)	45

---

---

12.8. Modèle 8 - k plus proches voisins régressif	45
12.9. Modèle 9 - Forêt aléatoire régressive	46
12.10. Modèle 10 - Forêt extra-aléatoire régressive	46
12.11. Modèle 11 - AdaBoost régressif	46
12.12. Modèle 12 - Gradient Boosting régressif	47
12.13. Modèle 13 - XGBoost régressif	47
12.14. Modèle 14 - LightGBM régressif	48
<b>13. Interprétation des modèles</b>	<b>49</b>
13.1. Le coefficient de détermination ( $R^2$ )	49
13.2. L'erreur quadratique moyenne de la racine (RMSE)	50
13.3. Optimisation des hyperparamètres	51
13.4. Courbes d'apprentissage	52
<b>14. Choix du modèle final</b>	<b>53</b>
<b>15. Application de restitution visuelle des résultats</b>	<b>55</b>
<b>Conclusion</b>	<b>58</b>
<b>Annexes</b>	<b>60</b>
Annexe 1 : Dictionnaire des variables	60
Annexe 2 : Représentation graphique des valeurs aberrantes	63
Annexe 3 : Asymétrie	64
Annexe 4 : Taux d'inflation aux Etats-Unis	65
Annexe 5 : Taux de criminalité aux Etats-Unis	65
Annexe 6 : Indice des prix des logements aux Etats-Unis	66
Annexe 7 : Matrice de corrélation entre les variables explicatives numériques et la variable cible	66
Annexe 8 : Résultat ANOVA	67
Annexe 9 : Importance des caractéristiques	68
<b>Rapport métier</b>	<b>71</b>
Quelques chiffres clés	71
Analyse de la distribution de quelques paramètres clés	71
Analyse des données externes	76
Résultat de la prédiction des prix des logements	79

---

---

## Introduction

Depuis plusieurs années, le marché de l'immobilier fluctue. Plusieurs facteurs entrent en jeu, chacun contribuant à l'évolution complexe du marché immobilier. En effet, nous pouvons retrouver le facteur de l'économie qui peut jouer une rôle important dans le marché de l'immobilier. Lorsque l'économie progresse, la demande de logements augmente généralement, ce qui peut impacter positivement les prix des logements. Et inversement, une baisse de l'économie peut provoquer une baisse des prix des logements. De plus, un autre facteur qui peut impacter les prix des logements est la localisation des logements. Les prix des logements en métropole seront plus élevés que les prix des logements se situant en pleine campagne. Puis, les caractéristiques du logement seront aussi déterminants pour le prix du logement. Comme vous l'aurez compris plusieurs facteurs peuvent influencer le prix des logements.

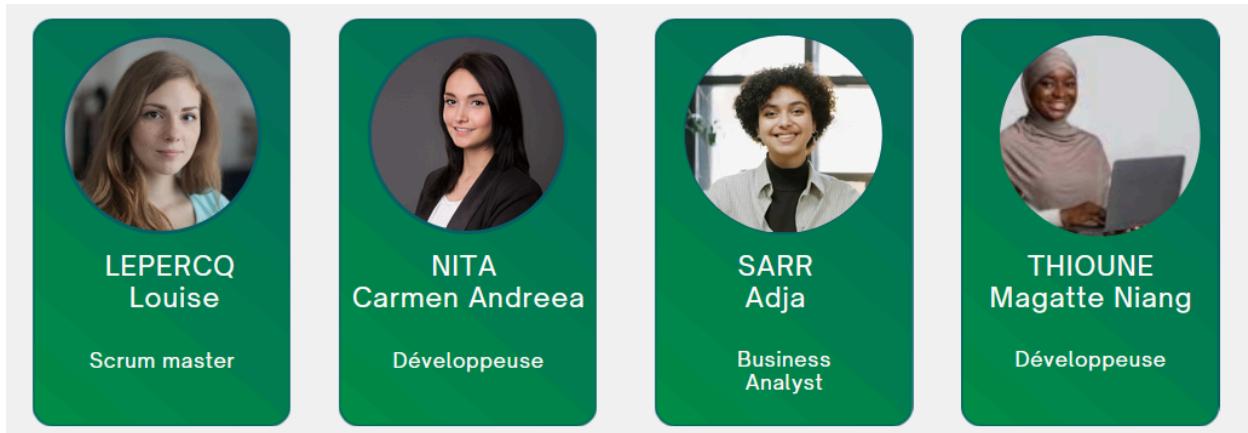
Aujourd'hui notre entreprise LNST Conseil a été missionnée de **réaliser une analyse prédictive des prix des logements**. Pour cela, nous avons un disposition un fichier dans lequel nous pouvons retrouver des caractéristiques du logement, ainsi que des informations sur la localisation du logement. Ce fichier va nous servir à tester différents modèles de prédiction et une fois affiner, le modèle sera capable de prédire le prix des logements selon leurs caractéristiques. Cette mission nous a été proposée par Kaggle et nous allons soumettre nos résultats afin d'obtenir le meilleur score possible et d'obtenir le meilleur classement.

Dans le cadre de ce rapport, nous avons établi un plan en plusieurs parties. Les premières parties seront consacrées à l'explication de notre organisation. Une partie présentera les données ainsi que la qualité de nos données. Une partie sera consacrée au feature engineering avec l'ajout de sources externes. Une partie consistera à connaître nos données à l'aide de statistiques descriptives. Une autre partie nous aidera à prendre des décisions sur le choix des variables que nous allons utiliser en partie pour la construction du modèle. Puis, nous allons finir par expliquer notre application, ainsi qu'une conclusion.

---

---

## 1. Présentation de l'équipe



Nous avons décidé d'effectuer le projet en mode agile, c'est-dire que chaque membre de l'équipe va réaliser ses tâches en autonomie. Cependant, s'il y a des points de blocages, tous les membres vont essayer de pallier les problèmes en trouvant des solutions. Pour un bon fonctionnement agile, le rôle de scrum master est important pour le bon fonctionnement de notre équipe. En effet, LEPERCQ Louise va organiser les tâches à effectuer lors de ce projet en les listant et en les attribuant aux membres de l'équipe.

Ensuite, NITA Carmen Andreea et THIOUNE Magatte Niang sont développeuses. Ce sont elles qui vont s'occuper de la partie développement de code, ce qui implique la conception, la programmation et les tests des fonctionnalités nécessaires pour résoudre la problématique du projet.

Puis SARR Adja a le rôle de Business Analyst. Son rôle est de comprendre le besoin métier et de les traduire de manière fonctionnelle. En d'autre terme, elle va s'occuper de la partie restitution pour avoir un visuel de nos résultats obtenus par les développements et participer à la rédaction du rapport.

Cependant, les rôles ne sont pas fixes puisque chaque membre de l'équipe essaie de partager ses connaissances et de passer un moment pour chaque étape du projet.

---

## 2. Mise en place de différent outils et organisation

### 2.1. Mise en place du Trello

Nous avons décidé de travailler à l'aide du “mode Agile”. Le mode agile est une approche de gestion de projet qui se caractérise par une flexibilité, une adaptabilité et une forte collaboration entre les membres de l'équipe. Les méthodes agiles peuvent prendre différentes formes : Scrum, Kanban, Extreme Programming (XP). Chacune de ces méthodes a ses propres pratiques spécifiques, mais toutes partagent l'objectif commun de favoriser la flexibilité et la réactivité face aux évolutions du projet.

Dans notre cas nous avons décidé Scrum et Kanban. En effet, nous avons utilisé une application appelée Trello qui permet de gérer notre projet. Nous allons donc créer des tâches et nous allons mettre à jour l'état de l'avancement de cette tâche.

Sur notre tableau de planification, nous avons décidé de créer quatres listes :

- 1) **A faire** : dans cette liste, nous allons lister les tâches que nous allons effectuer lors de ce projet ;
- 2) **En cours** : dans cette liste, nous allons glisser la tâche à faire dans la liste en cours lorsque nous allons commencer la tâche ;
- 3) **Selecture** : dans cette liste, nous allons glisser la tâche en cours lorsque nous considérons cette tâche terminée pour nous. Cependant, un autre membre de l'équipe va devoir relire notre travail afin qu'il soit réellement terminé ;
- 4) **Terminé** : dans cette liste, nous allons glisser les tâches que nous avons terminées.

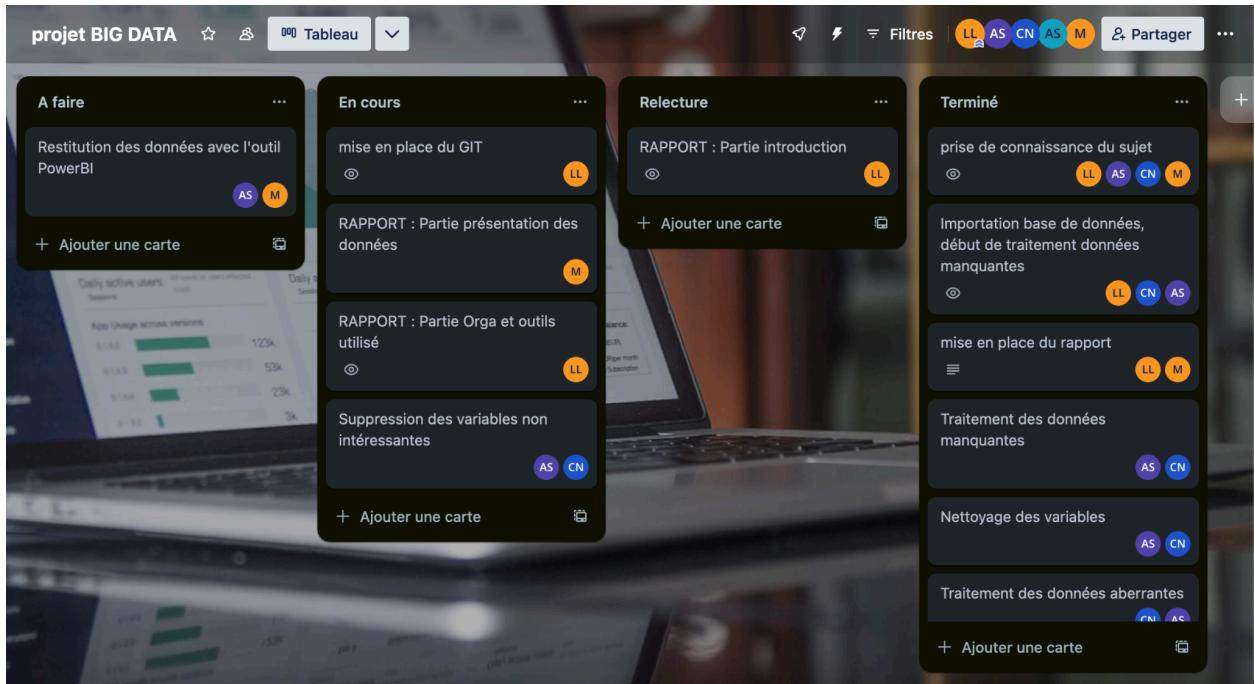


Figure 1 - Planification des tâches

Cette figure peut vous donner une idée plus précise sur notre mode d’organisation. Comme précisé auparavant, nous allons nous attribuer des tâches et nous allons mettre à jour le statut de notre tâche au fur et à mesure de l’avancement. Nous essayons également de mettre les tâches que nous allons faire dans le futur. Par exemple, nous avons déjà créé une tâche qui correspond à la restitution des données et a été attribuée à Magatte et à Adja. Cependant, il se peut que d’autres membres de l’équipe viennent épauler l’équipe pour cette tâche.

## 2.2. Mise en place du Git

Pour le bon fonctionnement de l’équipe, nous avons également mis en place un Git sur le GitLab de l’université. En effet, GitLab est une plateforme qui permet une gestion complète du cycle de vie des développements que nous effectuons. L’historisation des codes de développement nous permet de suivre l’évolution de notre projet au fil du temps, tandis que la possibilité de développer simultanément permet à chaque membre de l’équipe de travailler de

---

manière indépendante sur des fonctionnalités spécifiques. L'étape de fusion des versions à la fin du développement simplifie l'intégration des différentes contributions, assurant ainsi une cohérence globale du code. Cette approche collaborative renforce la productivité de l'équipe tout en facilitant la gestion des différentes branches de développement.

Notre repository	<a href="https://gitlab.univ-lille.fr/louise.lepercq.etu/lepercq_nita_sarr_thioune">https://gitlab.univ-lille.fr/louise.lepercq.etu/lepercq_nita_sarr_thioune</a>
Branche dev	Cette branche sera notre environnement de développement, c'est-à-dire que c'est sur cette branche que nous allons effectuer nos développements.
Branche main	Cette branche sera la branche principale du projet , c'est-à-dire qu'elle contiendra un code stable qui fonctionne sans erreur courante. Une fois que nous avons développé notre code dans la branche dev, et que nous nous sommes assuré du bon fonctionnement, nous pouvons basculer notre projet sur la branche main. Les deux branches fusionnent pour garder les informations déjà dans la branche main et pour ajouter les informations de la branche dev.

Tableau 1 - Information sur notre repository

Cette approche facilite le développement collaboratif et permet aux développeurs de travailler sur différentes parties du projet sans perturber la branche principale. La branche main sera donc réservée aux versions stables tandis que la branche dev sera réservée aux différents développements.

---

## 2.3. Organisation

Concernant notre organisation, nous avons décidé de découper le projet en différentes parties. En d'autres termes, nous pouvons travailler simultanément sur des étapes de notre projet. Nous allons décrire rapidement les différentes étapes que nous avons effectuées lors de ce projet. Nous allons donc revenir plus en précision dans la suite du rapport.

### ❖ **Étape 1 : 01\_Presentation\_nettoyage\_renommage**

Cette étape va permettre de connaître la base de données notamment le nombre de variables, leurs types. Elle va nous permettre également de repérer les données manquantes, aberrantes. Puis, elle va nous permettre de renommer les variables pour qu'elle soit plus compréhensive. Une fois ces étapes réalisées, nous allons extraire une nouvelle base de données que nous allons utiliser dans l'étape 2.

### ❖ **Étape 2 : 02\_donnees\_externes**

Cette étape va permettre des données externes, pour pouvoir croiser avec nos données initiales. Cette base complémentaire va nous aider à compléter nos données, ainsi que notre analyse. Nous allons ensuite faire une jointure avec nos différentes bases pour avoir une seule base de données avec toutes les informations. Nous allons donc utiliser cette nouvelle base de données dans l'étape 3.

### ❖ **Étape 3 : 03\_creation\_nouvelles\_variables**

Cette étape va avoir comme objectif de créer de nouvelles variables à partir d'une variable existante afin d'avoir de nouvelle information plus pertinente selon nous. Puis, dans cette étape, nous allons également faire des modifications sur les variables existantes, notamment

---

le changement d'unité. Pour la suite, nous allons utiliser cette nouvelle base de données pour l'étape 4.

❖ **Étape 4 : 04\_regroupements\_modalites**

Dans cette étape, nous allons réaliser des regroupements de modalités. Dans chaque analyse prédictive, il est important de faire cette étape car cela va permettre d'enlever des modalités rares, ou encore des modalités impertinentes. Cela va donc améliorer nos futurs modèles prédictifs. Nous allons donc utiliser cette nouvelle base de données dans l'étape 5 et l'étape 6.

❖ **Étape 5 : 05\_statistiques\_descriptives**

Dans cette étape, nous avons appris à connaître notre base de données car nous avons créé des graphiques illustrant certains aspects de notre base de données. Cette étape est importante pour mieux comprendre les liens potentiels entre les variables et connaître quelles variables peuvent faire varier le prix des logements.

❖ **Étape 6 : 06\_correlation\_variables**

Dans cette étape, nous allons vérifier les liens entre les variables car s'il en existe un, une redondance d'informations sera très possible dans notre future modélisation. Pour cette étape, nous avons repris la base que nous avons extraite en étape 4.

❖ **Étape 7 : 07\_modelisation**

Dans cette étape, nous avons réalisé la modélisation. C'est dans cette étape que nous allons tester différents modèles pour obtenir les meilleures prédictions possibles du prix des logements selon ses caractéristiques.

---

---

### 3. Présentation des données

#### 3.1. Description des données

L'ensemble de données utilisé dans cette étude se compose de 1460 enregistrements, chacun comportant 81 variables distinctes. Ces 81 variables explicatives décrivent (quasiment) tous les aspects des maisons résidentielles à Ames, dans l'État d'Iowa. Grâce à cette base de données nous allons pouvoir répondre à notre problématique qui est, nous le rappelons, de prédire le prix final de chaque logement. Pour avoir des précisions sur nos variables, nous avons établi un dictionnaire de variables fourni en annexe du document. Dans celui-ci, nous avons répertorié le nom de toutes les variables, le renommage des variables, leurs types, puis une brève description de chaque variable ([cf annexe 1](#)).

#### 3.2. Hypothèses

Dans cette sous-partie, nous allons établir quelques hypothèses sur les variables qui selon nous auront des impacts sur le prix du logement. Effectivement, il est important de s'interroger avant de faire des analyses sur ce qui pourrait jouer une rôle sur le prix du logement.

Impact sur le prix des logements	Explication
la superficie du logement	Nous pouvons penser que plus la maison est grande, plus le prix du logement sera élevé.
la localisation du logement	Nous pouvons penser que selon la localisation du logement, les prix de celui-ci peuvent fortement varier. Le prix des logements en centre ville seront peut-être plus élevés.
la qualité du logement	Nous pouvons penser que si la qualité du logement n'est pas optimale, il faut prévoir de potentiels travaux, donc le prix du logement serait peut être plus faible.
la présence de pièce annexe du logement	Nous pouvons penser que la présence d'un garage ou la présence d'un jardin peuvent influencer positivement sur le prix du logement.

Tableau 2 - Caractéristiques qui peuvent influencer le prix du logement

---

Émettre des hypothèses est une étape importante pour la suite de notre analyse. En effet, elles vont nous permettre de nous guider pour répondre à notre problématique qui nous le rappelons est de prédire les prix des logements se situant dans l'État d'Iowa.

Elles vont également nous aider à faciliter l'interprétation de nos résultats car nous pensons connaître quelques caractéristiques qui vont influencer les prix des logements. De plus, elles vont nous permettre de répondre à certaines questions des clients car s'ils nous demandent une maison à un certain prix, nous devons être en capacité de savoir quel type de maison pourrait lui correspondre, en lui donnant les caractéristiques du logement.

## 4. Mise en qualité des données

Avant de débuter notre analyse, il est impératif d'effectuer une étape cruciale de nettoyage des données. Au cours de cette exploration, nous avons repéré des données manquantes, caractérisées par une information inconnue, ainsi que des données aberrantes, ou anormales par rapport au reste de l'ensemble de données. Ces éléments pourraient compromettre la validité de l'interprétation statistique.

### 4.1. Valeurs manquantes

Une valeur manquante est une donnée qui n'est pas renseignée dans notre jeu de données. Cela peut se produire suite à différentes raisons, telles que la non connaissance de la valeur ou encore un oubli de saisie dans le jeu de données.

#### ❖ Détection des données manquantes :

Pour détecter les données manquantes, nous avons regardé pour chaque variable, s'il y avait la présence de valeur nulle. Dans notre jeu de données, ces données manquantes sont représentées par « NaN ». Le tableau ci-dessous révèle le nombre de valeurs manquantes pour différentes variables dans l'ensemble de données.

---

Variables	Nombre de valeurs	Variables	Nombre de valeurs
qualitePiscine	1453	conditionGarage	81
elementsDivers	1406	expositionSousSol	38
typeAlleeAcces	1369	qualiteSurfaceFinieSousSol2	38
cloture	1179	qualiteSousSol	37
qualiteCheminnee	690	conditionSousSol	37
longTerrainRue	259	qualiteSurfaceFinieSousSol1	37
typeGarage	81	typePlacageMaconnerie	8
anneeConstrGarage	81	superficiePlacageMaconnerie	8
interieurGarage	81	systElectrique	1

Tableau 3 - Valeurs manquantes

Certaines variables présentent un nombre significatif de valeurs manquantes, suggérant certaines tendances. Par exemple, la variable `qualitePiscine`, qui évalue la qualité de la piscine, et la variable `elementsDivers`, qui concerne des fonctionnalités diverses, affichent des proportions élevées de valeurs manquantes, indiquant probablement l'absence de piscine ou de caractéristiques spéciales. De même, les variables telles que `typeAlleeAcces`, `cloture` et `longTerrainRue` ont des valeurs manquantes qui pourraient signaler l'absence d'une allée spécifique, de clôture ou de données sur la distance par rapport à la rue.

Pour les variables liées au garage et au sous-sol, telles que `typeGarage`, `anneeConstrGarage`, `expositionSousSol` et `qualiteSurfaceFinieSousSol1`, les valeurs manquantes suggèrent l'absence respective de garage ou de sous-sol dans les maisons concernées. Les variables `typePlacageMaconnerie` et `superficiePlacageMaconnerie`, concernant le revêtement en pierre ou en briques, ont quant à elles un nombre limité de valeurs manquantes.. Enfin, la variable `systElectrique`, représentant le système électrique de la maison, a une seule valeur manquante.

---

#### ❖ Traitement des données manquantes :

Concernant le traitement des données manquantes, nous nous sommes rendu compte que ce n'est pas de vraies valeurs manquantes. En effet, pour la majorité des variables où nous repérons des valeurs manquantes, c'est le résultat de la non possessions de la caractéristique.

Pour les variables `anneeConstrGarage` et `superficiePlacageMaconnerie`, nous avons décidé de remplacer les valeurs manquantes par un 0. En effet, ces variables sont numériques et concerne une non possessions de la caractéristique. Nous avons fait le choix de mettre une modalité à 0 car pour la suite, nous risquerons d'être bloqué lors de potentiels regroupements. Si nous remplaçons par une modalité « inconnu », la variable sera transformée en format caractère (string), il sera donc impossible de comparer les valeurs en fonction d'une autre.

Pour la variable `longTerrainRue` qui est une variable numérique, nous avons décidé de remplacer les valeurs manquantes par la moyenne. Effectivement, les valeurs manquantes sont de vraies valeurs manquantes puisque cette variable représente

A propos de la variable `systElectrique` nous avons la présence d'une vraie donnée manquante. Nous avons donc remplacé cette valeur manquante par le mode.

Pour le reste des variables possédant des valeurs manquantes, nous avons donc décidé de créer une modalité « No + nom caractéristiques » pour la majorité des variables. Par exemple, pour la variable `qualitePiscine` nous avons créé une modalité « No PoolQc ».

## 4.2. Valeurs aberrantes

Une valeur aberrante est une valeur qui ne prend pas la même tendance que les autres. Elles peuvent être intégrées dans une base de données suite à différentes causes telles qu'une erreur de saisie, ou encore une erreur d'unité.

---

### ❖ Détection des données aberrantes :

Pour la détection de données aberrantes, nous avons réalisé des graphiques à boîte à moustache. Ce type de graphique permet de détecter facilement les données aberrantes pour chaque variable numérique. En effet, la boite à moustache permet d'avoir un visuel de distribution de nos données numériques. Elle se base sur des indicateurs statistiques tels que la médiane, le premier quartile, le troisième quartile puis l'étendu interquartile. Une donnée sera donc considérée comme aberrante si elle se trouve hors de la boîte à moustache, soit si elle n'appartient pas à l'étendu interquartile.

Cependant, la détection des données aberrantes est délicate car il se peut que la donnée existe vraiment. Effectivement, il peut y avoir des caractéristiques totalement différentes selon le logement, ce qui peut favoriser l'apparition excessive de données aberrantes. Donc nous avons considéré qu'il y a trois variables contenant des valeurs aberrantes.

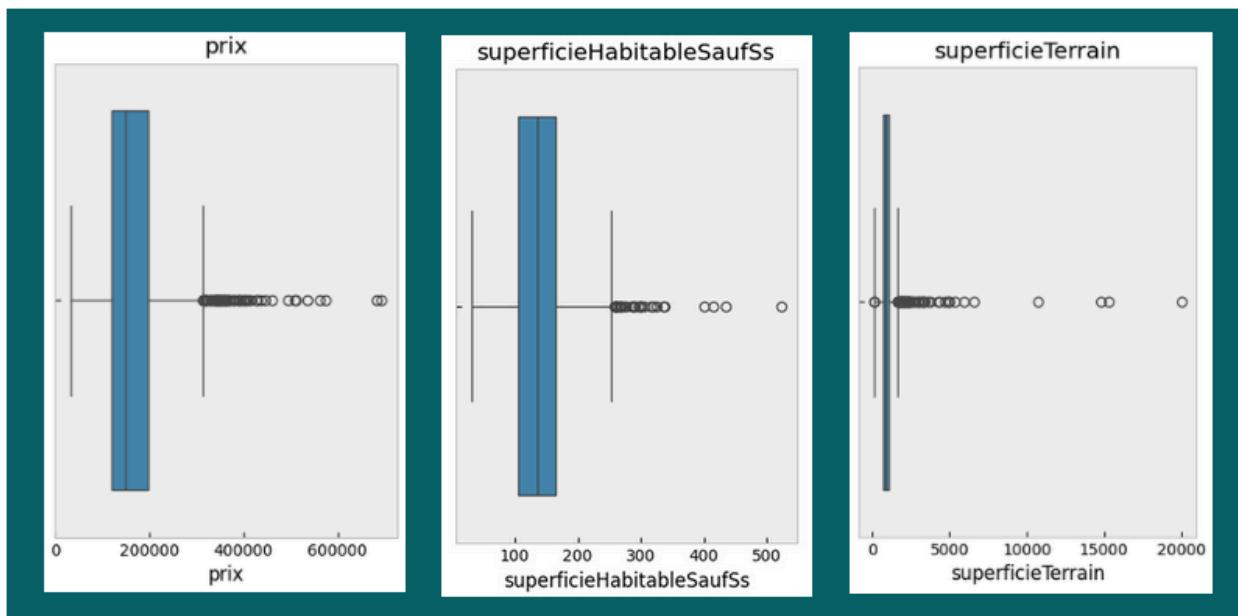


Figure 2 - Boîte à moustache contenant les valeurs aberrantes

*Remarque : une autre représentation est présente en annexe (cf [annexe 2](#))*

---

### ❖ Traitement des données aberrantes :

Comme nous l'avons dit précédemment, le traitement des données aberrantes est délicat car les caractéristiques du logement peuvent varier fortement. Donc, nous avons décidé de les traiter de manière très large et de supprimer peu d'observations.

Concernant la variable `prix`, nous avons opté pour la suppression des observations dont les prix excèdent le 99,5e percentile (0,995 quantile) de la distribution des prix. Cette approche semble judicieuse pour éliminer les valeurs aberrantes ou les points extrêmes qui pourraient potentiellement avoir un impact disproportionné sur la robustesse du modèle. Donc en d'autre terme, nous avons supprimer les lignes ayant un prix supérieur à \$ 527 332.

Concernant la variable `superficieHabitableSaufSs`, nous avons décidé de supprimer les observations ayant une superficie habitable hors sous sol supérieur à 4000 ft<sup>2</sup>.

Concernant la variable `superficieTerrain`, nous avons regardé dans un premier temps les observations ayant une superficie de terrain supérieur à 50 000 ft<sup>2</sup>. Cependant, nous avons remarqué que deux classes se démarquent, les logements aux alentours de 50 000 ft<sup>2</sup>, puis des logements aux alentours de 100 000 ft<sup>2</sup>. Donc nous avons décidé de supprimer les observations ayant plus de 100 000 ft<sup>2</sup>.

Après avoir parlé des données manquantes et aberrantes, nous allons maintenant parler des modifications des variables existantes, ainsi que des créations de variables.

## 5. Feature engineering<sup>1</sup>

### 5.1. Ajout de différentes sources externes

Pour enrichir nos analyses, nous avons décidé d'utiliser différentes sources externes. En effet, elles vont nous permettre de connaître si les prix des logements peuvent varier selon de

---

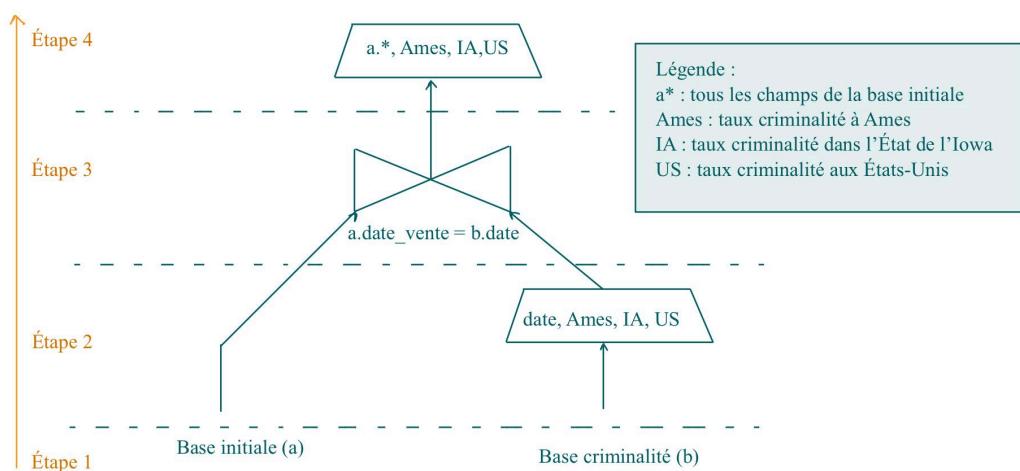
<sup>1</sup> Le feature engineering est un processus de préparation des variables afin qu'elles puissent être utilisées pour entraîner un modèle d'apprentissage automatique.

nouvelles informations que nous n'avons pas à disposition dans notre base de données initiale. Avant d'intégrer une source de données externes, nous avons d'abord réfléchi à ce que pourrait faire fluctuer le prix des logements, puis ensuite, nous pourrons potentiellement trouver l'information dans une base de données existante.

Nous nous sommes interrogés sur le fait que les logements se trouvent dans une ville des Etats-Unis (Ames) où le port d'armes est autorisé et un taux de criminalité élevé. Nous avons donc pensé que le taux de criminalité pourrait jouer un rôle sur les prix des logements.

Pour cela, nous avons trouvé une base de données<sup>2</sup> dans laquelle nous avons le taux de criminalité par an aux États-Unis, dans l'État de l'Iowa et dans la ville de Ames entre 2000 et 2017. Ces données sont issues des informations du Federal Bureau of Investigation (FBI). Ces trois taux vont nous permettre de pouvoir comparer et voir si le taux de criminalité est plus élevé dans la ville de Ames que dans les autres villes de l'État. Puis, par la suite, nous allons pouvoir si le prix du logement va augmenter ou diminuer selon le taux de criminalité. Pour pouvoir utiliser ces nouvelles données, nous avons réalisé une jointure entre notre base de données initiale et la base de données recensant le taux de criminalité. Nous avons donc réalisé cette jointure à l'aide d'une clé de jointure qui est de la date de vente du logement pour notre base initiale, et la date associée au taux de criminalité. Nous pouvons penser que le prix des logements va diminuer si le taux de criminalité est élevé dans certains quartiers.

Schéma 1 -  
Explication  
d'une jointure



<sup>2</sup> [base de données sur la criminalité](#)

---

Nous allons faire une grève explication de chaque étape de la jointure réaliser pour obtenir une nouvelle base avec de nouvelles informations :

- ❖ **étape 1** : Nous avons à disposition deux bases de données (notre base initiale que nous avons surnommé a, puis notre base criminalité que nous avons surnommé b) ;
- ❖ **étape 2** : nous avons sélectionné les champs de la base criminalité que nous voulons ajouté à notre base initiale ;
- ❖ **étape 3** : nous avons réaliser la jointure à l'aide de clé de jointure indiqué sur le schéma (date) ;
- ❖ **étape 4** : nous sélectionnons tous les champs que nous voulons en résultat. Dans notre cas, nous voulons tous les champs de la base initiale, ainsi que les champs sélectionnés en étape 2 de la base criminalité.

A la suite de cette jointure, nous allons donc obtenir une nouvelle base de données répertoriant tous les champs de la base initiale, ainsi que les nouveaux champs qui nous intéressent dans la base criminalité. Par la suite, nous allons à nouveau réaliser des jointure, et nous informons que nous allons prendre la base avec les nouveaux champs (soit celle après la jointure).

Ensuite, de nos jours en France l'inflation est un sujet redondant, donc nous nous sommes demandé si aux Etats-Unis c'était aussi le cas. Pour cela, nous avons opté pour une base de données<sup>3</sup> incluant le taux d'inflation par année aux États-Unis depuis 1960, ainsi que son évolution par rapport à l'année précédente. Ces données sont issues de World Bank. Comme pour la base recensant le taux de criminalité, nous avons eu recours à une jointure entre notre base initiale et la base ayant le taux d'inflation aux États-Unis par année. Nous avons utilisé les mêmes clés de jointure pour pouvoir joindre les deux bases de données, soit l'année de vente du logement pour notre base initiale et la date associée aux taux d'inflation pour notre nouvelle base de données sur l'inflation. Nous pouvons penser que si l'inflation sera élevée, le prix du logement sera aussi élevé.

---

<sup>3</sup> [base de données sur l'inflation](#)

---

De plus, nous avons trouvé une base de données<sup>4</sup> concernant l'indice de prix des logements dans la ville. C'est un indicateur qui mesure les variations des prix des logements dans la ville de Ames. Dans nos données, notre base 100 correspond à l'année 1995, cela signifie que les prix des logements en 1995 sont utilisés comme point de référence avec une valeur de 100. Toutes les données ultérieures sont alors ajustées par rapport à cette valeur de base. Par exemple, si en 1995 l'indice des prix des logements était de 100, et qu'en 2000 il était de 120, cela signifierait que les prix des logements à Ames ont augmenté en moyenne de 20 % par rapport à leur niveau en 1995.

Comme nos deux sources de données externes, nous allons effectuer une jointure entre notre base de données initiale et notre base de données sur l'indice de prix. Nous utilisons donc comme clé de jointure dans notre base initiale la date de vente des logements et pour notre nouvelle base, la date de l'indice de prix.

Pour finir avec les sources externes, nous nous sommes rendu compte que nous n'avions pas les coordonnées des quartiers que nous avons dans notre base initiale. Pour pouvoir réaliser une ou plusieurs visualisations sur la localisation des logements dans l'État de l'Iowa, nous allons créer une base de données répertoriant les noms de quartiers que nous avons dans notre base de données ainsi que leurs coordonnées associées. Pour cela, nous avons répertorié tous les quartiers que nous avons dans notre base de données, puis nous avons été sur Google Maps pour récupérer la latitude et la longitude du quartier.

## 5.2. Préparation de la base de données

### 5.2.1. Modifications de variables

Nous allons apporter des modifications sur certaines de nos variables. En effet, pour rendre plus compréhensif nos variables, nous avons décidé de changer d'unité, notamment pour les superficies, les longueurs, ou encore pour le prix du logement.

---

<sup>4</sup>[base de données sur l'indice de prix](#)

Variables	Unité de la variable	Unité de la variable converti
superficieTerrain superficieTotaleSousSol superficieHabitableSaufSs superficieGarage superficieTerrasseBois superficiePorcheOuvert	Pied carré (ft <sup>2</sup> )	Mètre carré (m <sup>2</sup> )
longTerrainRue	Pied (ft)	Mètre (m)
prix	Dollar (\$)	Euro (€)

Tableau 4 - Modification de variable

### 5.2.2. Crédation de variables

Dans notre étude, la prédiction des prix de vente des logements aux États-Unis nécessite la création de nouvelles variables qui est cruciale pour plusieurs raisons :

- ❖ les nouvelles variables peuvent fournir des informations supplémentaires qui ne sont pas directement disponibles dans les données d'origine. Ces informations peuvent aider à capturer des aspects importants des logements qui influent sur leurs prix de vente, ce qui améliore la capacité des modèles à faire des prédictions précises ;
- ❖ les relations entre les caractéristiques des logements et leurs prix de vente peuvent être complexes et non linéaires. En créant de nouvelles variables à partir des caractéristiques existantes, nous pourrons mieux représenter ces relations complexes et permettre aux modèles de capturer des motifs plus subtils dans les données ;

- 
- ❖ certaines caractéristiques des logements peuvent avoir plus d'impact sur les prix de vente lorsqu'elles sont combinées ou transformées d'une certaine manière. Par exemple, la proximité à une autoroute seule peut ne pas être très informative, mais la création d'une variable binaire indiquant si un logement est situé à proximité d'une autoroute ou non peut être plus pertinente pour prédire les prix de vente ;
  - ❖ les acheteurs peuvent juger important d'évaluer la durée d'existence d'une propriété avant de l'acheter. En effet, les logements plus modernes peuvent être vendus plus rapidement et à des prix plus abordables que des logements anciens. Raison pour laquelle les variables `ancienneteConstruction` et `ancienneteRenovation` ont été créées pour connaître respectivement la durée d'existence d'un logement et la durée de rénovation de ce logement (s'il a bien été rénové).

**Voici les hypothèses et explications des nouvelles variables créées :**

**Variable : `ancienneteRenovation`**

**Explication :** Cette variable mesure depuis combien d'années un logement a été rénové.

**Hypothèse :** Elle peut être importante car les logements récemment rénovés peuvent avoir une valeur différente de ceux qui ne l'ont pas été récemment. Les acheteurs peuvent être prêts à payer un prix plus élevé pour les logements rénovés récemment.

**Variable : `ancienneteConstruction`**

**Explication :** Cette variable indique le nombre d'années d'existence d'un logement depuis sa construction.

**Hypothèse :** Cela peut être pertinent car la date de construction peut influencer la qualité de la construction, les normes de construction en vigueur à l'époque et donc le prix de vente.

---

**Variable : superficieTotale**

**Explication :** Cette variable représente la superficie totale d'un logement en additionnant la superficie habitable et la superficie du sous-sol.

**Hypothèse :** La superficie totale peut être un facteur déterminant dans la détermination du prix de vente, car les acheteurs considèrent souvent la superficie totale de la propriété comme un critère important.

**Variable : mois\_anneeVente**

**Explication :** Cette variable concatène le mois et l'année de vente d'un logement.

**Hypothèse :** Elle peut être utilisée pour analyser les tendances saisonnières dans les prix de vente des logements et pour évaluer comment les prix varient au fil du temps.

**Variables : route\_ville, autoroute , proximite\_gare, proximite\_gare, proximite\_parc**

**Explication :** Ces variables binaires indiquent la proximité d'un logement à différents types d'infrastructures (route principale, autoroute, gare, parc).

**Hypothèse :** La proximité de ces infrastructures peut avoir un réel impact sur le prix de vente. Par exemple, une proximité avec une autoroute peut être considérée comme un inconvénient pour certains acheteurs en raison du bruit, tandis que la proximité d'un parc peut être perçue comme un avantage.

**Variables : nbTotalSallesBain**

**Explication :** Cette variable calcule le nombre de salles de bain dans le logement. Donc nous regroupons les salles de bains situées aux étages, ainsi que les salles de bains situées au sous-sol.

**Hypothèse :** Le nombre de salles de bain peut être un atout pour le prix du logement car les acheteurs ayant des enfants peuvent être prêts à payer plus cher leurs logements s'ils possèdent une famille nombreuse.

---

#### Variables : noteGlobale, noteMoyenne

**Explication :** Cette variable calcule la note globale donnée pour l'état du logement. Nous l'avons établie étant la somme ou la moyenne de la qualité du logement et de la condition du logement.

**Hypothèse :** L'état du logement sera un facteur important dans le prix du logement, donc il nous semblait important de créer une variable représentant cela.

Après la mise en qualité des données, l'ajout de champs grâce à des sources externes, ainsi la modification et création de champs à partir des champs que nous avions à disposition, nous allons pouvoir réaliser un visuel de nos champs les plus importants dans un modèle conceptuel de données.

## 6. MCD (Modèle Conceptuel de Données)

Un Modèle Conceptuel de Données (MCD) est un outil utilisé dans le domaine de la conception de bases de données pour représenter de manière abstraite la structure et les relations entre les données dans un système d'information. Le MCD ci-joint représente les données de notre système de gestion immobilière. Il décrit les entités principales du système, leurs attributs et les relations entre elles.

Le MCD suivant identifie les entités suivantes :

- ❖ **Garage** : Stocke des informations sur les garages, y compris leur type, leur condition, l'année de construction du garage ;
  
- ❖ **Surface** : Stocke des informations sur les surfaces, y compris la superficie totale, la superficie du terrain, la superficie du garage, la superficie de la piscine.

- 
- ❖ **Qualité Logement** : Stocke des informations sur la qualité du logement, y compris la qualité globale, la qualité du sous-sol, la qualité de la cuisine, la qualité de la piscine, le type du bâtiment, le style du bâtiment, le nombre d'années de construction et de rénovation d'un logement ;
  - ❖ **Localisation** : Stocke des informations sur la localisation du bien immobilier, y compris le quartier, la proximité de l'autoroute, de la route principale, de la gare et du parc ;
  - ❖ **Pièce** : Stocke des informations sur les pièces du bien immobilier, y compris le nombre de pièces, le nombre de salles de bain, le nombre de demi-salles de bain, le nombre de chambres, le nombre de cuisines, le nombre de cheminées et le nombre de places de voitures ;
  - ❖ **Vente (T)** : Stocke des informations sur les ventes de biens immobiliers, y compris le mois et l'année de la vente, le type de vente, la condition de vente, le prix de chaque logement.

Chaque entité (c'est-à-dire chaque table) possède un ensemble d'attributs qui décrivent ses caractéristiques. Le MCD présenté offre une vue d'ensemble des données du système de gestion immobilière. Il permet de comprendre les différentes entités du système, leurs attributs et les relations entre elles (voir figure 3).

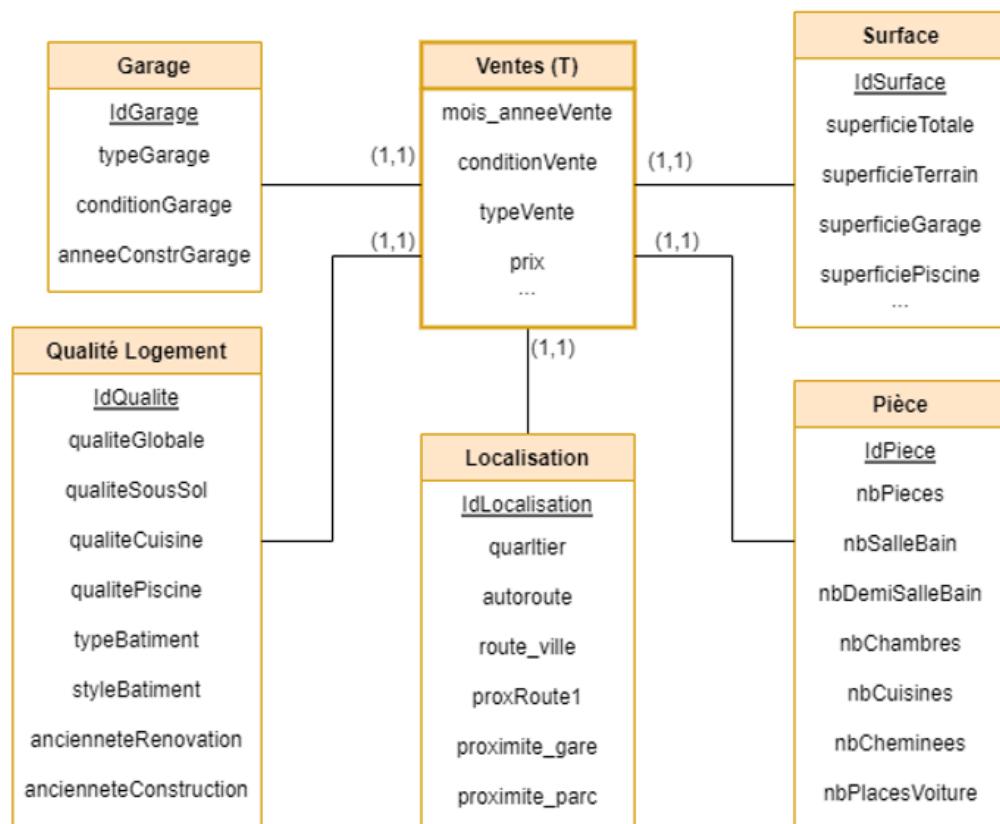


Figure 3 - Modèle conceptuel de données

## 7. Statistiques descriptives

Les statistiques descriptives fournissent une base solide pour toute analyse statistique ultérieure. Elles aident à comprendre la nature des données, à détecter des problèmes potentiels et à orienter les choix méthodologiques dans le cadre de l'étude sur les prix de logements. Pour rappel, après tous les traitements de certaines variables (création des variables, suppression de variables, modifications de variables), nous allons réaliser nos statistiques descriptives à l'aide d'une base de données avec **106 variables** et 1460 observations.

---

## 7.1. Statistiques univariées

Dans cette partie de l'étude, notre attention se portera exclusivement sur l'analyse et la description des caractéristiques spécifiques d'une variable à la fois. Ce processus permet de mettre en lumière la distribution, la tendance centrale et la dispersion des données, fournissant ainsi une vision détaillée des propriétés intrinsèques de la variable sans tenir compte de ses relations avec d'autres variables.

### ❖ Variable cible : Prix des logements

Pour connaître notre variable cible qui est le prix des logements dans la ville de Ames dans l'État Iowa aux États-Unis, nous avons calculé quelques indicateurs statistiques de base pour nous donner une idée sur les prix des logements de notre base de données.

**Prix minimum**

**32 610.34 €**

**Prix moyen**

**166 490.15 €**

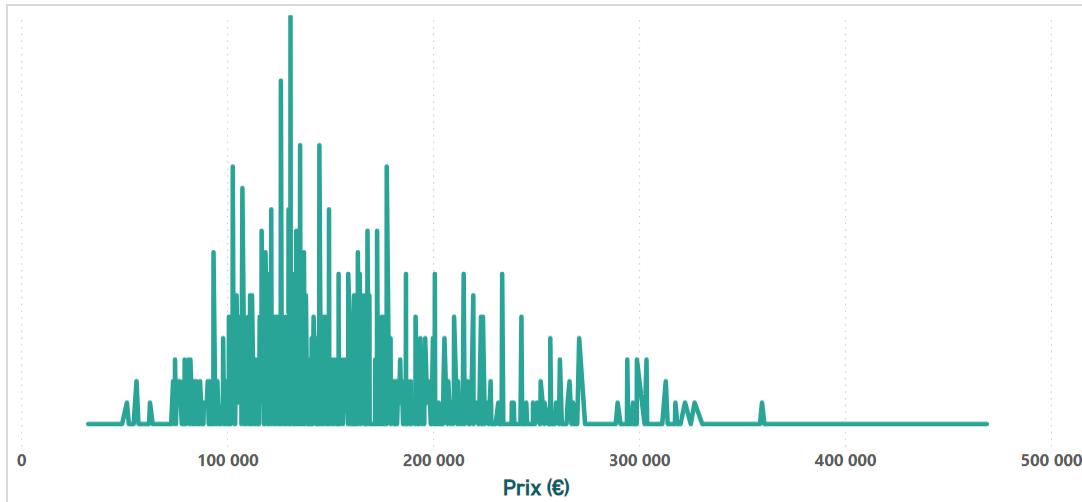
**Prix maximum**

**468 913.37 €**

De plus, pour voir un peu mieux la distribution des prix des logements, nous avons réalisé un graphique représentant la distribution des prix. L'objectif de visualiser la distribution des prix est de visualiser la répartition des prix des logements pour connaître les tendances ou encore la dispersion des prix. Nous pourrons donc identifier s'il y a des valeurs extrêmes. Puis, nous pouvons également étudier l'asymétrie<sup>5</sup> de la distribution. En effet, cela pourrait fournir des informations sur la nature des données et influencer le choix des techniques utilisées (notamment pour le remplacement d'éventuelles données manquantes).

---

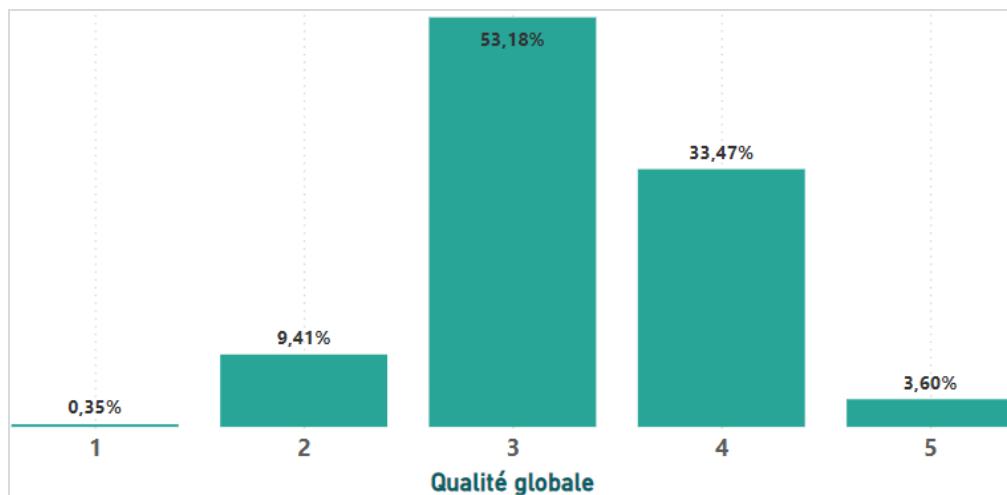
<sup>5</sup> L'asymétrie sera expliquée en [annexe 3](#)



Graphique 1 - Distribution des prix des logements

Le graphique précédent montre que la majorité des logements ont des prix situés dans la fourchette de 100 000€ à 200 000€. La distribution des prix des logements est assez variée, avec une tendance à la concentration autour des prix moyens, mais avec également une présence significative de logements aux prix plus élevés. Cela suggère une diversité dans l'offre de logements, avec des options disponibles pour différents budgets et préférences.

#### ❖ Distribution des logements (en %) en fonction de la qualité globale



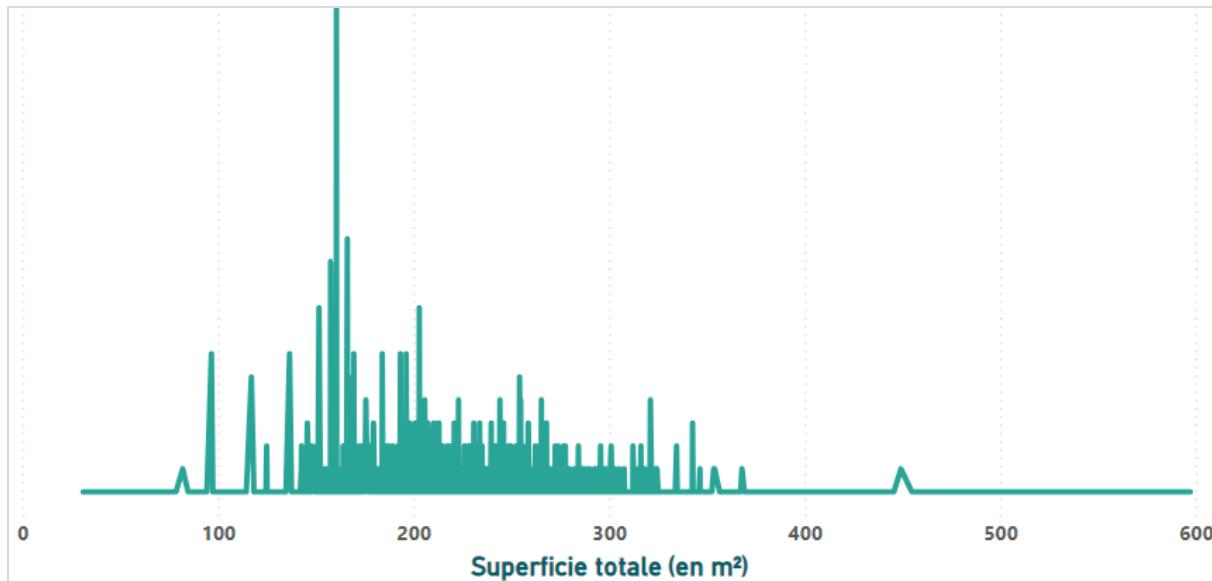
Graphique 2 - Distribution de la qualité globale des logements

---

Le graphique montre la distribution de la qualité globale des logements de la ville de Ames. La qualité est définie sur une échelle de 1 à 5, où 1 et 2 représentent des qualités faibles ; 3 et 4 des qualités moyennes et au-delà de 5 il s'agit des logements avec de très bonnes qualités. Nous remarquons que les logements avec une faible qualité sont plus nombreux que les logements avec une qualité moyenne. Les logements de haute qualité sont moins fréquents.

Le marché immobilier examiné se compose principalement de maisons de qualité modeste à moyenne. Il n'y a pas de maisons haut de gamme, ce qui pourrait être attribuable aux prix.

#### ❖ Distribution des logements en fonction de la superficie totale



Graphique 3 - Distribution de la superficie totale des logements

Le graphique montre la distribution des superficies totales en mètres carrés des logements. Dans la ville d'Ames, Iowa, aux États-Unis, la répartition de la surface totale des logements présente une asymétrie vers la droite. La majorité des logements aux États-Unis ont une superficie comprise entre 100 et 350 mètres carrés.

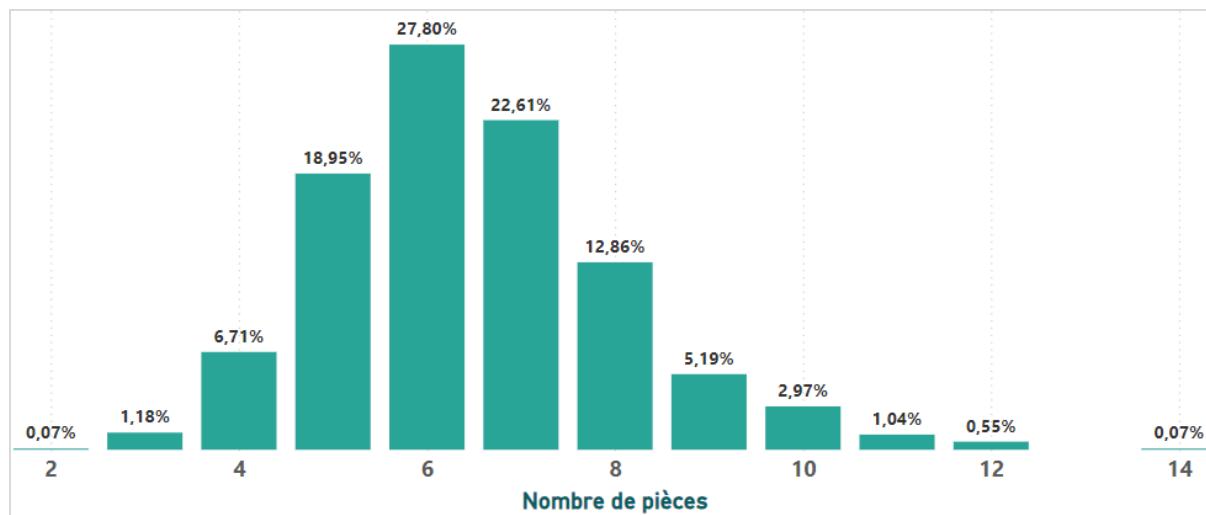
---

La prédominance des logements de superficie comprise entre 100 et 350 mètres carrés peut s'expliquer par plusieurs facteurs. Ces types de logements correspondent aux besoins de familles de tailles différentes. Ils offrent un bon équilibre entre l'espace et le prix. Ils sont généralement plus faciles à trouver et à vendre que les logements avec une superficie atypique.

La proportion plus faible de logements de petite superficie (moins de 100 mètres carrés) peut s'expliquer par le fait que ces logements sont souvent trop petits pour la plupart des familles aux Etats-Unis La proportion plus faible de logements de grande superficie (plus de 300 mètres carrés) peut s'expliquer par le fait que ces logements sont généralement plus chers et moins adaptés aux familles de taille moyenne.

Les logements de plus petite superficie (moins de 100 mètres carrés) et de plus grande superficie (plus de 350 mètres carrés) sont minoritaires.

#### ❖ Distribution des logements (en %) selon le nombre de pièces



Graphique 4 - Distribution du nombre de pièces du logement

Le graphique montre la répartition du nombre de pièces dans les logements de la ville de Ames. La majorité des logements de cette ville (environ 69%) ont un nombre de pièces compris entre 5 et 7 pièces. Les logements avec 8 pièces sont également assez présents, représentant

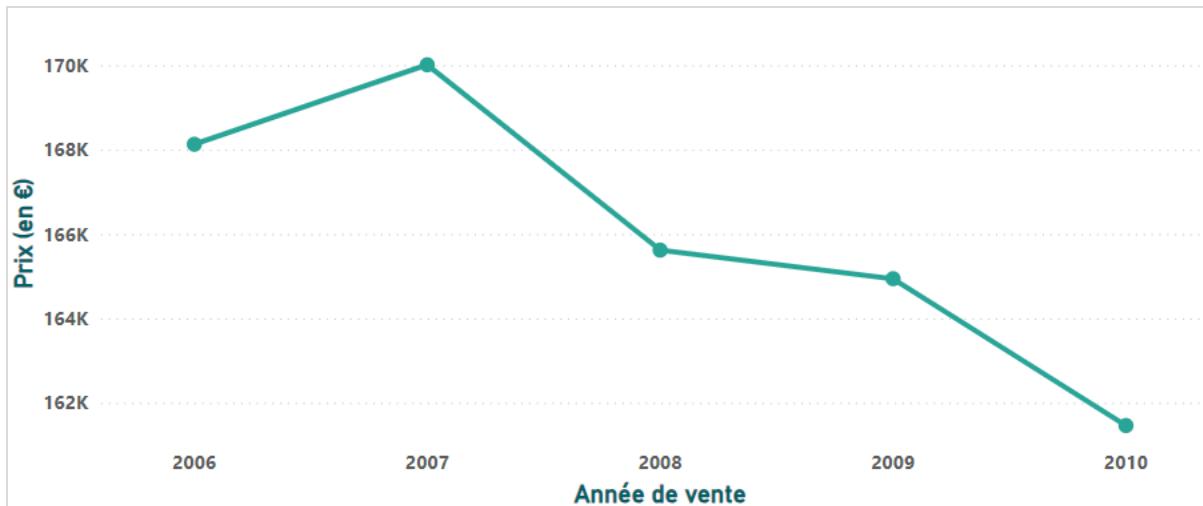
environ 13% de la répartition totale. Les logements ayant moins de 5 pièces (environ 8%) et plus de 8 pièces (environ 10%) sont moins fréquents.

La prédominance des logements avec 5, 6 ou 7 pièces peut s'expliquer par le fait que ces types de logements correspondent plus aux familles de moyennes ou de grandes tailles, d'où le nombre de ventes important par rapport aux autres logements avec moins de pièces ou plus.

## 7.2. Statistiques bivariées

Dans cette section, notre attention sera portée sur l'étude simultanée de deux variables, avec pour objectif de saisir les relations, corrélations et interactions qui peuvent émerger entre elles. Notre démarche consistera à explorer le lien entre la variable cible, à savoir le prix des logements, et d'autres variables que nous considérons comme pertinentes dans cette étude.

### ❖ Distribution des prix en fonction de l'année de vente



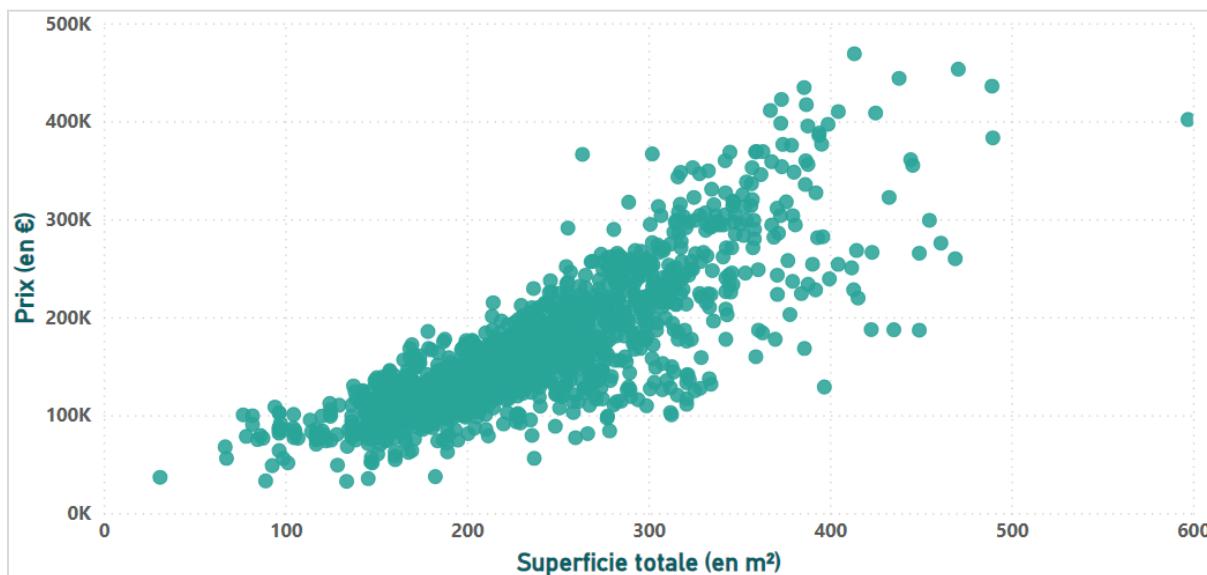
Graphique 5 - Distribution des prix en fonction de l'année de vente

Le graphique montre une variation des prix des biens immobiliers vendus entre 2006 et 2010. Nous remarquons que le prix des logements a légèrement augmenté entre 2006 et 2007 allant de 168 000 € à 170 000 €. Ensuite, une baisse importante est constatée en 2008. Cette

diminution plus importante des prix entre 2007 et 2008 peut être due à un événement particulier, tel qu'une crise économique ou une catastrophe naturelle. Enfin, le prix diminue à nouveau progressivement jusqu'en 2010 en passant de 165 800 € à 162 000 €.

Ces variations de prix peuvent s'expliquer par plusieurs facteurs, dont les caractéristiques du logement, l'augmentation ou la diminution de la demande, l'augmentation ou la diminution des coûts de production ou l'inflation.

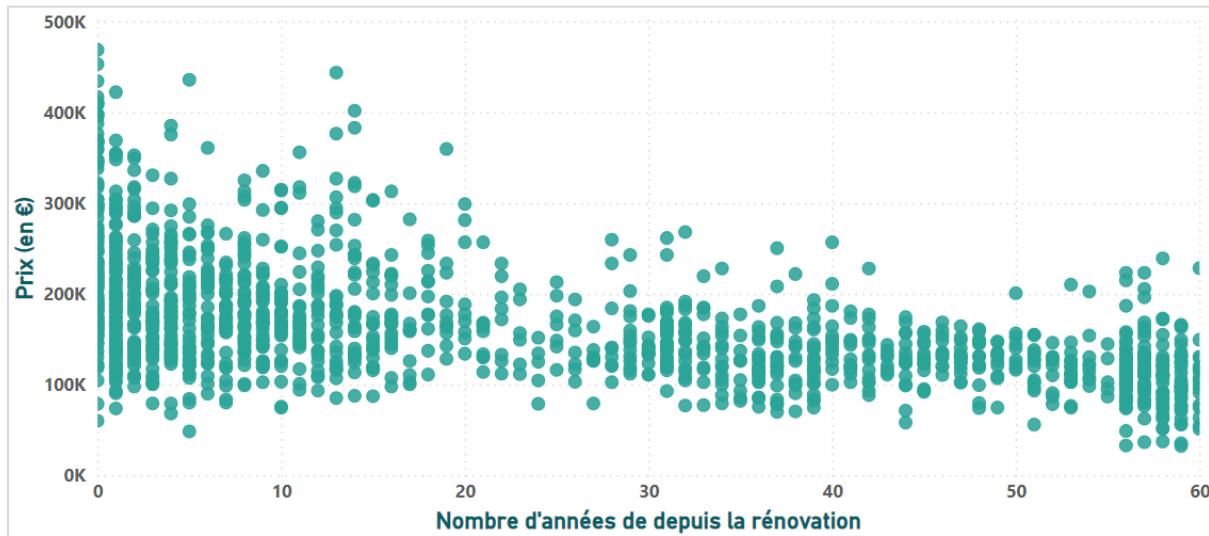
❖ **Distribution des prix en fonction de la superficie totale des logements :**



Graphique 6 - Distribution des prix en fonction de la superficie totale des logements

Une relation positive est observée entre la surface totale et le prix des logements, indiquant que généralement, une augmentation de la surface totale s'accompagne d'une hausse des prix. De plus, une dispersion plus étendue des données est constatée dans la tranche de prix la plus élevée, ce qui suggère une variation plus importante des prix parmi les logements de grande taille par rapport à ceux de petite taille.

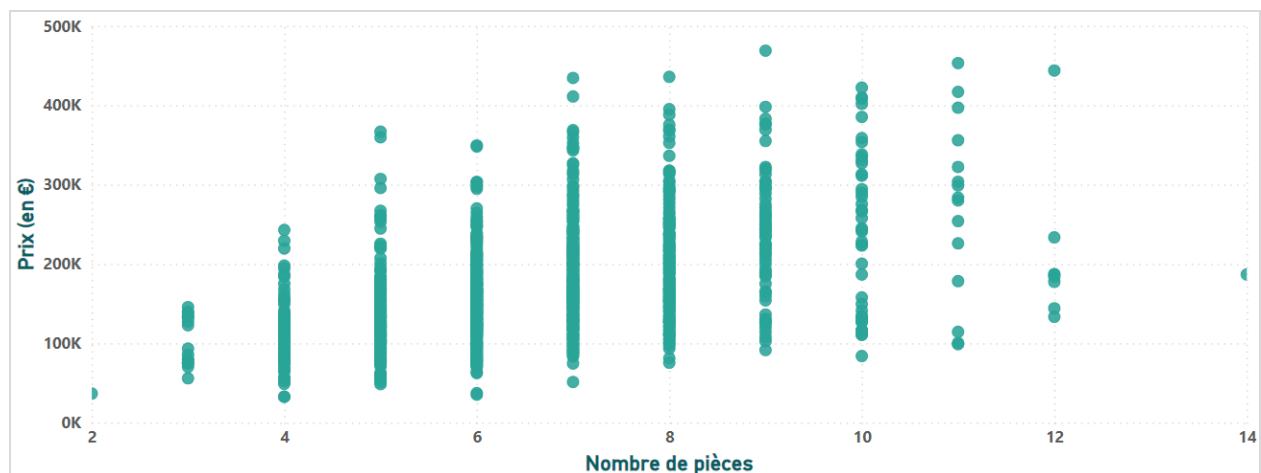
❖ Distribution des prix en fonction du nombre d'années depuis la rénovation :



Graphique 7 - Distribution des prix en fonction du nombre d'années depuis la rénovation

Le coût des logements décroît à mesure que leur rénovation s'éloigne dans le temps. Mais la baisse de prix est plus prononcée pour les logements rénovés depuis plus de 30 ans. Les logements neufs sont donc les plus chers.

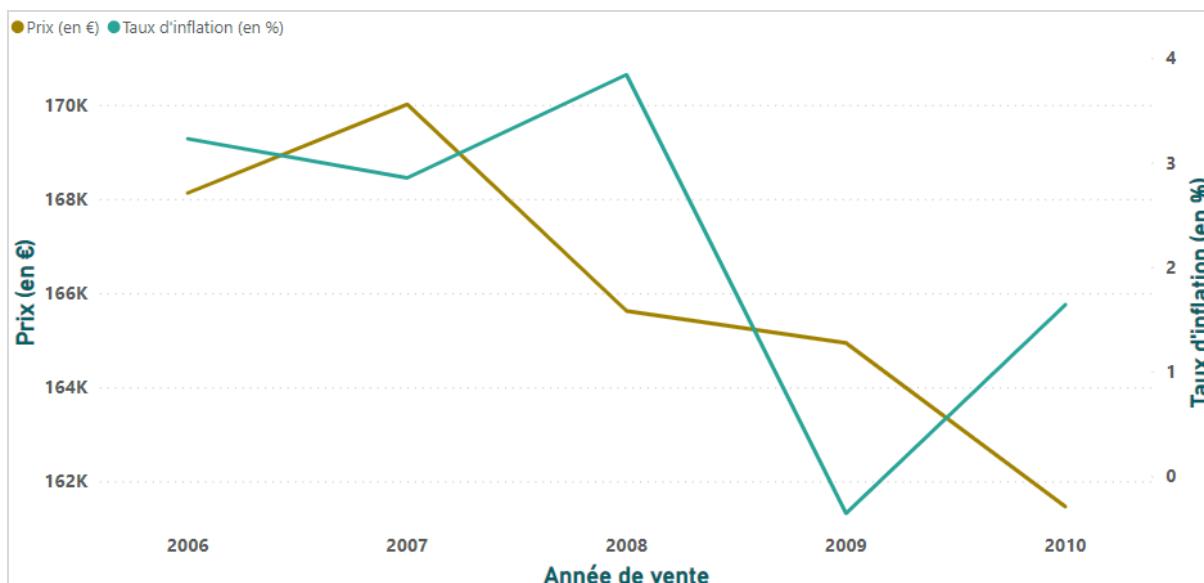
❖ Distribution des prix en fonction du nombre de pièces des logements :



Graphique 8 - Distribution des prix en fonction du nombre de pièces des logements

Nous constatons que le prix augmente à mesure que le nombre de pièces augmente, ce qui indique une relation positive entre les deux. Cependant, cette augmentation de prix n'est pas constante. Elle est lente au début, puis s'accélère entre 4 et 8 pièces, avant de se stabiliser à nouveau.

#### ❖ Distribution des prix et du taux d'inflation en fonction de l'année de vente

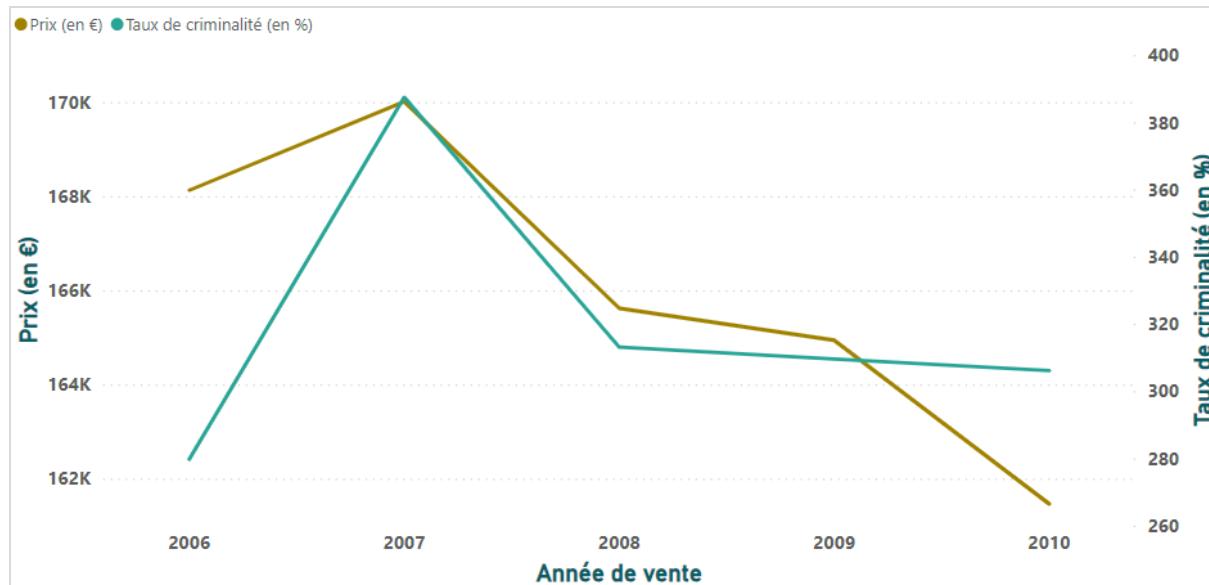


Graphique 9 - Distribution des prix et du taux d'inflation en fonction de l'année de vente

Nous remarquons que le prix et le taux d'inflation sont proportionnels. Lorsque le prix augmente, le taux d'inflation diminue. De même lorsque le prix diminue, le taux d'inflation augmente. Une baisse importante du taux d'inflation et des prix est constatée en 2008.

Cette période correspond à la crise financière mondiale qui a été une crise économique majeure débutée aux États-Unis et s'est ensuite propagée au reste du monde. La crise a été causée par un certain nombre de facteurs, notamment la bulle immobilière aux États-Unis et la crise des prêts hypothécaires à risque. La crise a eu un impact important sur l'économie mondiale, provoquant une récession, une augmentation du chômage et une baisse des prix des actifs. Nous avons mis en [annexe 4](#) une autre représentation graphique réalisée avec un autre logiciel.

❖ **Distribution des prix et du taux de criminalité dans la ville d'Âmes en fonction de l'année de vente**

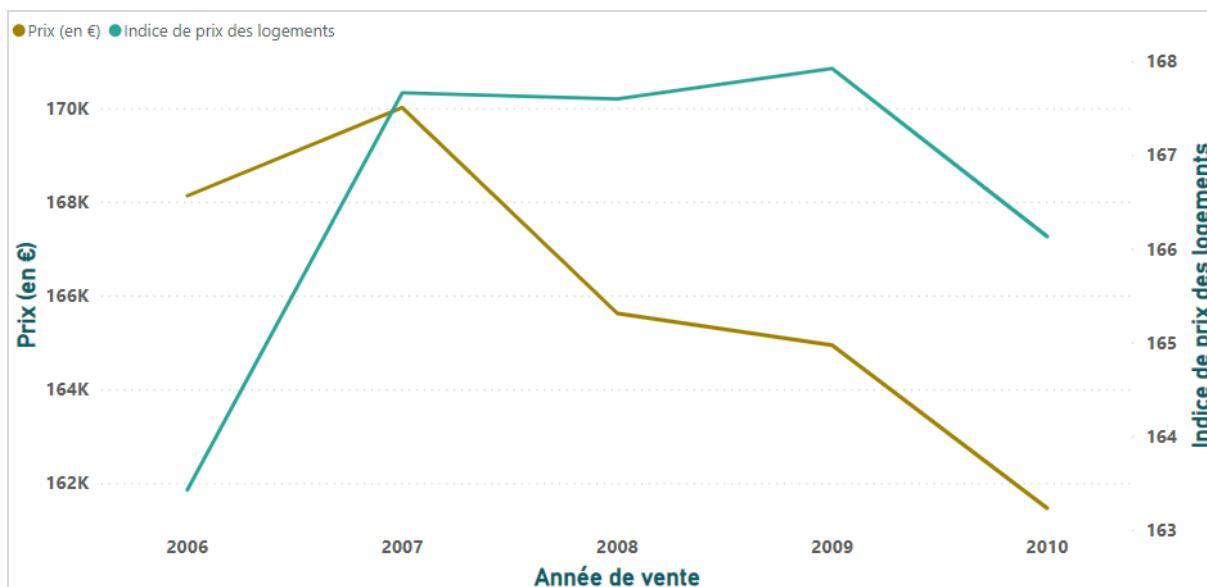


Graphique 10 - Distribution des prix et du taux de criminalité en fonction de l'année de vente

Le prix et le taux de criminalité ne semblent pas suivre une tendance générale similaire. Nous notons des variations importantes des prix des logements et du taux de criminalité d'une année à l'autre. En 2008, nous constatons une décroissance rapide de la distribution des deux variables. Cette baisse importante est certainement due à la crise économique mondiale qui a largement impacté les activités aux Etats-Unis notamment le secteur immobilier.

Le graphique montre une absence de corrélation directe entre le prix et le taux de criminalité dans la ville d'Ames. Il est possible que d'autres facteurs influencent le taux de criminalité, tels que le niveau de pauvreté, le taux de chômage ou la présence de *gangs*. Nous avons mis en [annexe 5](#) une autre représentation graphique réalisée avec un autre logiciel.

## ❖ Distribution des prix et de l'indice de prix en fonction de l'année de vente



Graphique 11 - Distribution des prix et du taux de criminalité en fonction de l'année de vente

Le graphique précédent montre la distribution des prix de vente des logements aux USA et de l'indice de prix en fonction de l'année de vente (entre 2006 et 2010). Entre 2006 et 2007, la distribution des prix de vente des logements et de l'indice de prix a progressivement augmenté : le prix passe de 168 000 € à 170 000 € et l'indice de prix augmente de 68 environ par rapport à l'indice de base 100 en 1995.

Le graphique montre que le prix et l'indice de prix des logements aux USA ont connu une tendance à la baisse entre 2007 et 2008 surtout la distribution des prix, probablement due à la crise financière mondiale de 2008 provoquant ainsi un effondrement du marché immobilier américain. Nous avons mis en [annexe 6](#) une autre représentation graphique réalisée avec un autre logiciel.

---

## 8. Pertinence des variables

Au vue des statistiques descriptives, nous pouvons penser que certaines variables ne seront pas pertinentes pour la modélisation. Pour cela, nous avons décidé de supprimer 13 variables qui nous ont servi principalement pour les statistiques descriptives, mais qui n'ont pas d'intérêt à être intégré dans le modèle. Donc à présent, pour la suite de l'analyse, nous allons avoir une base de données de 93 variables.

id	date_taux_inflation	Longitude
mois_anneeVente	annee_taux_inflation	Latitude
ancienneteConstruction	date_taux_criminalite	Coordonnées
ancienneteRenovation	annee_taux_criminalite	
changement_annuel_taux_inflation	annee_indice_prix_logements	

Tableau 5 - Variables à supprimer

## 9. Relation entre les variables

Il est important d'étudier la relation entre les différentes variables que nous avons à disposition. Pour cela, nous allons utiliser plusieurs pratiques tel que le coefficient de corrélation entre deux variables qualitatives, la procédure anova entre une variable quantitative et une variable qualitative.

---

## 9.1. Coefficient de corrélation

Le coefficient de corrélation est une mesure statistique permettant de connaître les potentielles relations entre deux variables qualitatives. En effet, ce coefficient va déterminer à quel point la variation d'une variable va influencer la variation d'une autre. Concernant l'interprétation du coefficient de corrélation :

- ❖ plus le coefficient est proche de 1, plus les variables sont liées positivement entre elles ;
- ❖ plus le coefficient est proche de -1, plus les variables sont liées négativement entre elles ;
- ❖ plus le coefficient est proche de 0, plus les variables n'ont pas de lien entre elles.

Pour la meilleure interprétation, nous allons définir différents seuils que nous allons respecter. Nous allons considérer qu'une variable catégorielle sera liée avec la variable cible lorsque leurs coefficients de corrélation seront supérieur à 0,3 quand elles seront corrélées positivement et -0,3 quand elles seront corrélées négativement.

Puis, nous allons considérer qu'une variable catégorielle sera liée avec une autre variable catégorielle, lorsque leurs coefficients de corrélation seront compris entre 0,6 et 1 lorsqu'elles seront liées positivement et entre -0,6 et -1 lorsqu'elles seront liées négativement. Dans ce cas, nous allons supprimer l'une des deux variables pour éviter la redondance d'information, afin d'éviter une mauvaise influence dans notre futur modèle. Nous avons obtenu un visuel de la matrice de corrélation que nous allons mettre en [annexe 7](#).

---

## 9.2. Procédure ANOVA

L'ANOVA, ou l'analyse de la variance est une technique statistique utilisée afin d'examiner l'influence de plusieurs variables catégorielles sur une variable numérique. Le test évalue si la variation entre les moyennes des groupes est statistiquement significative, c'est-à-dire si elle ne peut pas simplement être due au hasard. Si la p-valeur associée au test est inférieure au

seuil de significativité de 0.05, l'hypothèse nulle est rejetée en faveur de l'hypothèse alternative, indiquant qu'il y a des différences significatives entre au moins deux des moyennes de groupe.

En d'autres termes, l'ANOVA nous permet de comprendre si la variable explicative influence la variable cible et comment. Si la p-valeur est faible, nous avons des preuves statistiques pour suggérer que les moyennes des groupes sont différentes, et nous pouvons conclure que les variables catégorielles ont une certaine influence sur la variable continue Y (prix).

<b>hypothèse nulle</b>	Les moyennes des groupes définis par les différentes catégories de (X) sont égales. En d'autres mots, il n'y a aucune différence significative dans les moyennes des groupes.
<b>hypothèse alternative</b>	Au moins une paire de moyennes des groupes définis par les différentes catégories de (X) est significativement différente. Il existe donc une différence significative dans au moins une paire de moyennes des groupes.

Tableau 6 - Hypothèses de la procédure ANOVA

Nous avons donc réalisé à l'aide de python la procédure anova, puis nous avons obtenu pour chaque variable la statistique anova (F), la p\_value (P), puis sa signification. Le résultat (cf [annexe 8](#)) nous a permis de savoir si nous devons rejeter ou non l'hypothèse selon les variables, et ainsi de savoir si les variables sont significatives ou non.

En examinant les résultats de l'ANOVA pour les différentes variables catégorielles, nous pouvons tirer plusieurs observations significatives. En général, lorsque la valeur p est inférieure à 0.05, nous rejetons l'hypothèse nulle, ce qui suggère que les moyennes des groupes sont statistiquement différentes. Prenons les exemples de deux variables : 'zonage' et 'utilites'.

Variables	Statistique ANOVA F Valeur	P	Signification
zonage	80.036903	1.094003e-33	Rejeter
utilites	0.318766	5.724378e-01	Ne pas rejeter

Tableau 7 - Exemple de sortie de la procédure ANOVA

---

D'après le tableau précédent, la variable `zonage` possède un F élevé et un p très proche de 0, ce qui suggère que les moyennes des groupes définis par les différentes zones sont statistiquement différentes. Ce résultat implique que nous devons rejeter l'hypothèse nulle, nous devons donc prendre en compte l'hypothèse alternative. Pour rappel l'hypothèse alternative est que la variable est significativement différentes de celles des autres.

A l'inverse, la variable `utilites` possède un F faible et un p supérieure à 0,05, ce qui indique qu'il n'y a pas suffisamment de preuves pour rejeter l'hypothèse nulle. Cela suggère que les moyennes des groupes définis par les différents types d'utilitaires ne sont pas statistiquement différentes. Par conséquent, nous ne pouvons pas conclure que le type d'utilitaires a un impact significatif sur les prix des biens immobiliers en fonction de ces résultats d'ANOVA.

Les résultats indiquent que ces variables catégorielles spécifiques ont un impact significatif sur les prix des biens immobiliers, car les différences entre les moyennes des groupes sont statistiquement importantes, sauf pour les variables `utilites`, `penteTerrain`, `qualitePiscine`, `contourTerrain`, `typeRouteAcces`, `contourTerrain`, `proxRoute2`, `chauffage`, `fonctionnalites`, `elementsDivers`, `route_ville`, `proximite_gare` et `proximite_parc`. Par conséquent, nous envisageons de supprimer ces variables de notre étude, car elles ne semblent pas contribuer de manière significative à la variation des prix des biens immobiliers. Cependant, avant de prendre cette décision définitive, d'autres facteurs sont à prendre en compte, tels que la pertinence conceptuelle, le contexte et la logique métier pour éviter de baser exclusivement notre décision sur des critères statistiques.

Après analyse de la répartition des modalités de chacune de ces variables ainsi que leurs sens métier, nous avons décidé de supprimer ces variables `utilites`, `penteTerrain`, `qualitePiscine`, `contourTerrain`, `fonctionnalites`, `route_ville`, `proxRoute2`.

Après avoir étudié les liens entre les variables, nous allons continuer notre analyse à l'aide de 76 variables.

---

## 10. Préparation des données pour le Machine Learning

Depuis le début, nous travaillons sur une base de données train qui permet d'entraîner notre modèle. Cela signifie que notre base de données contenait des variables explicatives et une variable cible, soit le prix. Notre modèle doit donc s'entraîner sur cette base de données afin d'étudier et d'apprendre les relations que peuvent avoir les différentes variables et la variable cible. Cela va donc permettre d'obtenir une meilleure estimation du prix du logement à l'aide des différentes variables explicatives. Après toutes les modifications sur cette base train, nous avons une base de données de 2903 observations et de 75 variables. Pour un meilleur fonctionnement des modèles, nous allons encoder nos variables, c'est-à-dire que pour chaque modalité, elles seront transformées en numérique.

Variable	Modalités	Modalités encodées
qualitePiscine	'None'	0
	'Fa'	1
	'Gd'	2
	'Ex'	3

Tableau 8 - Exemple d'encodage

Ensuite, nous avons également eu une base test ayant 1490 observations et 80 variables. Cette base comprend les mêmes variables que la base train avait initialement à l'exception du prix. Nous allons traiter la qualité des données de manière similaire que celle pour la base train. En effet, nous allons traiter les données manquantes de la même manière (avec les mêmes règles de remplacement) que pour les données manquantes de la base train. Cependant, pour la base test, nous n'allons pas traiter les données aberrantes car pour le bien du concours Kaggle, nous devons avoir absolument le même nombre d'observations au début et à la fin de notre analyse (1459). De plus, nous avons effectué les mêmes modifications que nous avons effectué sur la

---

base train, soit l'imputation des données manquantes, la création de variables et la modifications de variables.

En d'autre terme, nous avons dû réaliser des changements sur la base de données test pour que sa composition soit la même que la base train. Le but de cela est que le modèle se sera entraîné avec une base ayant certaines caractéristiques, donc si la base test ne comporte pas les mêmes caractéristiques, il ne sera pas prédire les prix pour la base test.

## 11. Problème rencontré pour le Machine Learning

Nous avons donc utilisé ces variables pour réaliser différents modèles. Nous obtenons des résultats assez satisfaisant, cependant le classement sur Kaggle n'était pas très bon puisque nous étions classés parmi les derniers. Nous avons essayé de comprendre ce résultat et nous nous sommes rendu compte que nos sources externes jouent un rôle défavorable dans notre classement Kaggle. Effectivement, Kaggle calcul un score en fonction des prédictions des prix, et le fait que le prix prenait en compte aussi des informations qui influencent le prix, notre classement n'était pas le reflet de la réalité. De plus, nous nous sommes également rendu compte que nos changements d'unité pour le prix, ainsi que pour les surfaces jouaient un rôle négatif pour notre classement Kaggle.

Au vu de nos résultats sur Kaggle, nous avons donc décidé de ne pas prendre en compte nos sources externes dans les modèles. Donc pour la suite des constructions des modèles, nous allons utiliser une base de données de 81 variables, celle de la base train initiale. Nous gardons cependant nos traitements pour la qualité des données (données manquantes, données aberrantes), puis nous allons également garder les conversions des unités.

---

## 12. Construction et évaluation des modèles

### 12.1. Modèle 1 - Régression linéaire

La régression linéaire est une technique statistique fondamentale utilisée pour modéliser la relation entre une variable dépendante continue et une ou plusieurs variables indépendantes, appelées variables explicatives. L'objectif est de trouver la meilleure droite (ou plan, dans le cas de plusieurs variables explicatives) qui minimise la différence entre les valeurs observées et prédites de la variable dépendante, mesurée par des écarts appelés résidus. Cette droite est caractérisée par une équation linéaire de la forme :

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \varepsilon,$$

où  $Y$  est la variable dépendante,  $\beta_0$  est l'ordonnée à l'origine,  $x_i$  représentent les variables explicatives,  $\beta_i$  sont les coefficients de régression à estimer, et  $\varepsilon$  est le terme d'erreur. En ajustant les coefficients de manière à minimiser les résidus, la régression linéaire permet de modéliser et de prédire les relations linéaires entre les variables.

### 12.2. Modèle 2 - Lasso

La régression Lasso, également connue sous le nom de Least Absolute Shrinkage and Selection Operator, est une méthode de régression linéaire qui ajoute une pénalité L1 à la fonction de perte pour régulariser les coefficients du modèle. Elle cherche à minimiser la somme des moindres carrés des écarts entre les prédictions du modèle et les valeurs réelles de la cible, tout en pénalisant les coefficients de régression pour éviter le surajustement. La pénalité L1 favorise la sparsité en encourageant certains coefficients à devenir exactement nuls, ce qui peut entraîner la sélection automatique des variables. La force de la régularisation est contrôlée par un paramètre de régularisation, généralement noté alpha. La régression Lasso est souvent utilisée pour la sélection de variables et la réduction de dimensionnalité dans les problèmes de régression

---

où un grand nombre de caractéristiques sont présentes, car elle peut aider à identifier les variables les plus importantes pour la prédiction.

### 12.3. Modèle 3 - Ridge

Ridge Regression, également connue sous le nom de régression de crête, est une méthode de régression linéaire qui ajoute une pénalité L2 à la fonction de perte pour régulariser les coefficients du modèle. L'objectif est de minimiser la somme des moindres carrés des écarts entre les prédictions du modèle et les valeurs réelles de la cible, tout en pénalisant les coefficients de régression pour éviter le surajustement. La force de la régularisation est contrôlée par un paramètre de régularisation, généralement noté alpha. Ridge Regression est efficace pour traiter les problèmes de régression où les caractéristiques sont fortement corrélées ou lorsque le nombre de caractéristiques est proche de la taille de l'échantillon. Elle peut aider à stabiliser les estimations des coefficients du modèle et à améliorer la généralisation en réduisant la variance des estimations.

### 12.4. Modèle 4 - Kernel Ridge

Kernel Ridge Regression est une méthode de régression qui étend la régression ridge en introduisant une fonction de noyau pour gérer des données non linéaires. Tout comme la régression ridge, elle cherche à minimiser la somme des moindres carrés des écarts entre les prédictions du modèle et les valeurs réelles de la cible, tout en pénalisant les coefficients de régression pour éviter le surajustement. Cependant, au lieu d'utiliser une fonction de noyau, elle utilise une matrice de noyau pour transformer les caractéristiques d'entrée dans un espace de dimension supérieure où les données peuvent être linéairement séparables. Cela permet de modéliser des relations plus complexes entre les variables d'entrée et la variable cible. Kernel Ridge Regression est particulièrement efficace pour traiter des ensembles de données où les relations entre les caractéristiques et la cible sont non linéaires, bien qu'elle puisse être sensible au choix du noyau et aux paramètres de régularisation.

---

## 12.5. Modèle 5 - ElasticNet

ElasticNet est une technique de régression régularisée utilisée en apprentissage automatique pour réduire le surajustement et améliorer la généralisation des modèles linéaires. Cette méthode combine à la fois les pénalités L1 (lasso) et L2 (ridge) dans la fonction de perte, permettant ainsi une sélection automatique des variables et une régularisation des coefficients du modèle. Le paramètre alpha contrôle le mélange des pénalités L1 et L2, tandis que le paramètre de régularisation lambda contrôle la force de la régularisation. ElasticNet est particulièrement utile lorsque les données présentent des corrélations entre les caractéristiques et lorsque le nombre de caractéristiques est élevé par rapport à la taille de l'échantillon. En ajustant les paramètres alpha et lambda, ElasticNet permet de trouver un compromis entre la réduction de la variance et le biais du modèle, offrant ainsi une meilleure performance prédictive.

## 12.6. Modèle 6 - Arbre de décision régressif

Un arbre de décision régressif (*Decision Tree Regressor*) est un modèle d'apprentissage automatique utilisé pour la régression, où l'objectif est de prédire une valeur continue plutôt que de classer des observations dans des catégories. L'arbre de décision divise récursivement l'espace des caractéristiques en sous-ensembles plus petits, en choisissant à chaque étape la caractéristique et la valeur de seuil qui maximisent la réduction de l'erreur de prédiction, généralement mesurée par la variance ou l'erreur quadratique moyenne. L'arbre est construit jusqu'à ce qu'un critère d'arrêt soit atteint, par exemple lorsque le nombre maximum de niveaux de l'arbre est atteint ou lorsque chaque feuille contient un nombre minimum d'observations. Lors de la prédiction, une observation est passée à travers l'arbre, et sa valeur cible est estimée en prenant la moyenne des valeurs des observations présentes dans la feuille correspondante. Les arbres de décision régressifs sont simples à interpréter et à visualiser, mais peuvent souffrir de surajustement lorsqu'ils sont trop profonds ou lorsqu'ils ne sont pas régularisés. Ils sont souvent utilisés en tant que composants de modèles plus complexes, tels que les forêts aléatoires ou les boosters.

---

## 12.7. Modèle 7 - Machine à Vecteurs de Support (SVM)

La Machine à Vecteurs de Support (SVR) à noyaux est une méthode d'apprentissage automatique utilisée à la fois pour la classification et la régression. Son principe fondamental est de trouver un hyperplan dans un espace de grande dimension qui sépare au mieux les différentes classes (ou qui approxime au mieux les valeurs cibles dans le cas de la régression). Lorsque les données ne sont pas linéairement séparables dans l'espace d'origine, les SVM à noyaux utilisent une technique appelée "noyau" pour projeter les données dans un espace de dimension supérieure où elles peuvent être séparées linéairement. Les noyaux les plus couramment utilisés incluent le noyau linéaire, le noyau polynomial et le noyau gaussien (RBF). Les SVM à noyaux sont particulièrement adaptées aux ensembles de données de taille moyenne à grande, offrant souvent une bonne généralisation et une résistance au surajustement, bien que leur entraînement puisse être relativement coûteux en termes de temps computationnel.

## 12.8. Modèle 8 - k plus proches voisins régressif

Le modèle des k plus proches voisins régressif (KNeighborsRegressor) est une technique d'apprentissage automatique non paramétrique utilisée pour la régression. L'idée principale est de prédire la valeur d'une nouvelle observation en prenant la moyenne des valeurs des k observations les plus proches dans l'espace des caractéristiques. La proximité est généralement mesurée en utilisant une distance comme la distance euclidienne. Ce modèle ne nécessite pas d'entraînement explicite ; il stocke simplement les données d'entraînement pour les utiliser lors de la prédiction. La performance du modèle dépend fortement du choix de k et de la mesure de distance utilisée. Le modèle des k plus proches voisins régressif est simple à mettre en œuvre, mais peut être sensible à la présence de valeurs aberrantes et au choix du nombre de voisins.

---

## 12.9. Modèle 9 - Forêt aléatoire régressive

La Forêt Aléatoire Régressive (*Random Forest Regressor*) est un modèle d'apprentissage automatique qui utilise un ensemble d'arbres de décision pour effectuer des prédictions de régression. Chaque arbre de décision est entraîné sur un sous-ensemble aléatoire des données d'entraînement et des caractéristiques, ce qui favorise la diversité des arbres. Lorsqu'une prédiction est nécessaire, chaque arbre de décision donne une estimation, et la forêt aléatoire prend la moyenne de ces estimations pour produire une prédiction finale. Cette approche permet de réduire le surajustement et d'offrir une meilleure généralisation par rapport à un seul arbre de décision. La Forêt Aléatoire Régressive est largement utilisée pour prédire des variables continues dans divers domaines, offrant une bonne précision et une robustesse face aux données bruitées ou aux valeurs aberrantes.

## 12.10. Modèle 10 - Forêt extra-aléatoire régressive

La Forêt Extra-Aléatoire Régressive (*ExtraTreesRegressor*) est une méthode d'apprentissage automatique qui appartient à la famille des arbres de décision ensemblistes. Contrairement à la forêt aléatoire traditionnelle, elle choisit aléatoirement les seuils de division pour les nœuds de l'arbre, en plus de sélectionner un sous-ensemble aléatoire de fonctionnalités à chaque étape de construction de l'arbre. Cette approche de sélection aléatoire permet de réduire davantage la variance et le surajustement, tout en offrant une meilleure généralisation. En agrégeant les prédictions de multiples arbres de décision construits de cette manière, la Forêt Extra-Aléatoire Régressive fournit une estimation robuste et précise des valeurs cibles dans les tâches de régression, adaptée à une variété de problèmes de prédiction.

## 12.11. Modèle 11 - AdaBoost régressif

AdaBoost régressif (*AdaBoostRegressor*) est une technique d'apprentissage automatique qui construit un modèle de régression en combinant plusieurs modèles de régression faibles de

---

manière séquentielle. À chaque étape, le modèle accorde plus de poids aux exemples mal prédits par les modèles précédents, ce qui permet de se concentrer sur les erreurs résiduelles. Les modèles faibles sont ensuite pondérés en fonction de leur performance dans la prédition de ces exemples. En combinant ces modèles, AdaBoost régressif produit un modèle fort capable de s'adapter à des relations complexes entre les variables d'entrée et la variable cible. Ce processus de pondération séquentielle permet à AdaBoost de s'améliorer itérativement, offrant ainsi une précision accrue dans les prédictions de régression.

## 12.12. Modèle 12 - Gradient Boosting régressif

Le Gradient Boosting régressif (`GradientBoostingRegressor`) est une méthode d'apprentissage automatique qui construit un modèle prédictif en combinant plusieurs modèles simples, généralement des arbres de décision, de manière itérative. À chaque étape, le modèle tente de corriger les erreurs résiduelles du modèle précédent en se concentrant sur les gradients de la fonction de perte par rapport aux prédictions actuelles. Cela permet au modèle d'ajuster progressivement ses prédictions pour minimiser l'erreur globale. Le Gradient Boosting régressif est largement utilisé pour résoudre des problèmes de régression, offrant une grande précision et une bonne capacité à modéliser des relations complexes entre les variables d'entrée et la variable cible.

## 12.13. Modèle 13 - XGBoost régressif

XGBoost (`XGBRegressor`) est une technique d'apprentissage automatique qui appartient à la famille des méthodes ensemblistes. Il est basé sur l'agrégation de multiples modèles de prédition simples, appelés arbres de décision, afin de construire un modèle plus puissant et plus précis. XGBoost utilise un processus itératif pour construire les arbres de décision, en mettant l'accent sur les erreurs des prédictions précédentes. Grâce à cette approche, XGBoost est particulièrement efficace pour modéliser des relations complexes et non linéaires entre les variables explicatives et la variable cible, et il est souvent utilisé pour des tâches de régression et

---

---

de classification. Son efficacité, sa flexibilité et ses performances en font un choix populaire dans de nombreux domaines d'application.

## 12.14. Modèle 14 - LightGBM régressif

Le modèle LightGBM (LGBMRegressor) est une technique d'apprentissage automatique particulièrement puissante pour la prédiction des prix immobiliers. LightGBM est une implémentation efficace et rapide du gradient boosting framework, qui construit un modèle de prédiction en agrégeant des arbres de décision. Contrairement à la régression linéaire, LightGBM est capable de capturer des relations non linéaires et complexes entre les variables explicatives et la variable cible, ce qui en fait un choix privilégié pour les tâches de prédiction de prix immobiliers.

En ajustant les paramètres du modèle pour minimiser une fonction de perte spécifique, LightGBM apprend à prédire les prix des maisons en utilisant des caractéristiques telles que la taille de la maison, le quartier, le nombre de chambres, etc. Son efficacité et sa capacité à modéliser des relations complexes en font un outil précieux pour maximiser les performances prédictives dans ce type de compétition.

Noms des modèles	Abréviations	Noms des modèles	Abréviations
Régression linéaire	LR	K plus proche voisin	KNN
Lasso	LSO	Forêt aléatoire régressive	RF
Ridge	RIDGE	Extra forêt aléatoire	ET
Kernel Ridge	KR	AdaBoost régressif	AB
ElasticNet	ELNT	Gradient Boosting régressif	GB
Arbre de décision	DT	XGBoost régressif	XGB
SVM	SVM	LightGBM régressif	LGB

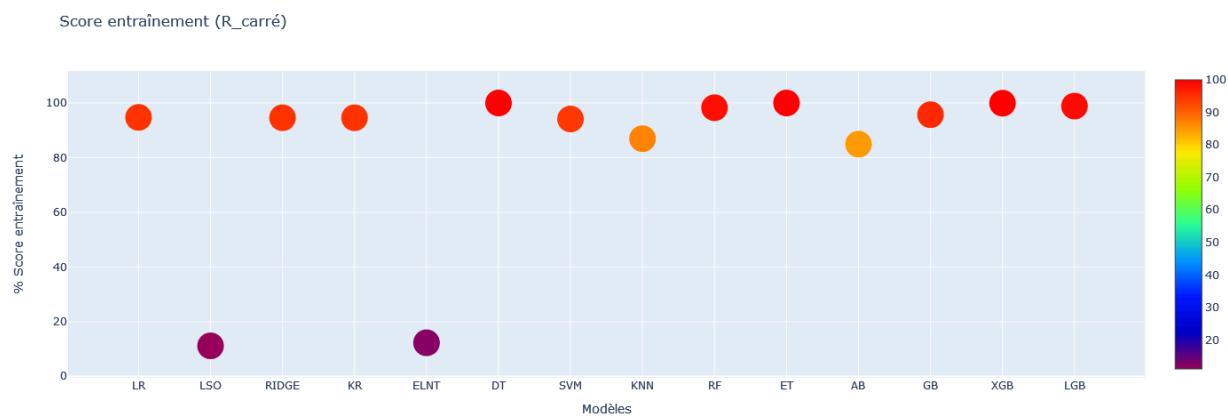
Tableau 9 - Abréviaison des noms de modèle

---

## 13. Interprétation des modèles

### 13.1. Le coefficient de détermination ( $R^2$ )

Le coefficient de détermination, ou aussi appelé  $R^2$  est un indicateur qui permet d'illustrer la qualité d'ajustement de nos modèles statistiques. Il mesure plus précisément la proportion de la variance de la variable dépendante qui est expliquée par le modèle. Concernant son interprétation, plus le  $R^2$  est proche de 1, plus le modèle est bon car il explique bien cette variance de la variable dépendante. A l'inverse, plus le  $R^2$  est proche de 0, moins le modèle est bon. Cependant, il ne faut pas oublier que le  $R^2$  n'est pas le seul indicateur pour évaluer la performance d'un modèle car un  $R^2$  correct peut seulement dire que le modèle s'ajuste bien au modèle. Donc pour chaque modèle nous avons étudié le  $R^2$  de chaque modèle pour avoir une idée du modèle qui s'ajuste le mieux aux données d'entraînement.



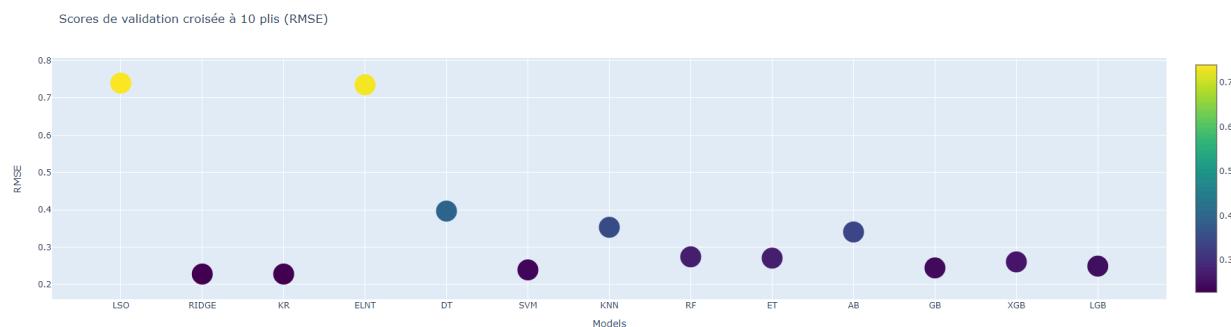
Graphique 12 -  $R^2$  des modèles

Comme nous pouvons le remarquer sur le graphique précédent, nous avons deux modèles qui possèdent un  $R^2$  nettement plus faible (environ de 11%, soit 0,1) que les autres, soit les modèles lasso et ElasticNet. Cependant, parmi les modèles qui possèdent un  $R^2$  parfait (à 100%, soit 1) sont ceux des modèles arbre de décision et extra forêt aléatoire. Comme précisé précédemment, ce n'est pas parce que le  $R^2$  est parfait que c'est ces modèles qui font le plus

fiable et le plus proche de la réalité. De plus, les autres modèles qui ont un  $R^2$  variant entre 84% et proche de 100% sont peut être aussi performant que ceux qui ont un  $R^2$  à 100%.

### 13.2. L'erreur quadratique moyenne de la racine (RMSE)

Comme le  $R^2$ , l'erreur quadratique moyenne de la racine aussi appelée RMSE, est un indicateur permettant d'évaluer la performance de nos modèles. Le RMSE mesure la différence entre les valeurs prédites par nos modèles et les valeurs réelles dans les données d'entraînement. Elle calcule donc l'écart quadratique moyen entre les prédictions du modèle et les valeurs réelles, puis prend la racine carrée de cette moyenne. Concernant l'interprétation, plus le RMSE est proche de 0, plus le modèle est proche de la valeur réelle. A l'inverse, plus le RMSE est proche de 1, plus le modèle est éloigné de la valeur réelle.



Graphique 13 - RMSE des modèles

Les RMSE des modèles du graphique précédent sont cohérent avec les résultats du  $R^2$ . En effet, les modèles Lasso et ElasticNet ont un RMSE proche de 0.8, donc ces deux modèles sont par définition les moins bons parmi les 14 modèles proposés. A l'inverse, les modèles ayant les meilleurs  $R^2$  n'ont pas le meilleur RMSE tel que l'arbre de décision.

### 13.3. Optimisation des hyperparamètres

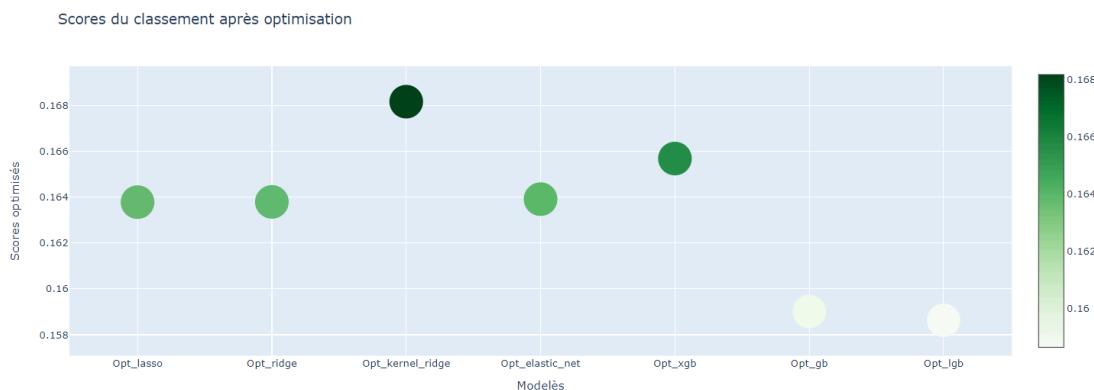
Au vu des résultats des  $R^2$  et des RMSE, nous avons décidé de sélectionner 8 modèles afin de les optimiser tels que :

Noms des modèles			
Lasso	Ridge	Kernel Ridge	ElasticNet
XGBoost régressif	Gradient Boosting régressif	LightGBM régressif	

Tableau 10 - Liste des modèles que nous allons optimisé

Pour améliorer nos modèles, nous avons décidé de faire varier les hypermètres de chaque modèle. Le but de cette étape est de rendre encore plus performant les modèles afin d'obtenir les meilleures prédictions possibles et aussi de mieux gérer le sur apprentissage du modèle. Après plusieurs variations des paramètres, nous allons garder les meilleurs modèles.

Nous allons ensuite relancer nos modèles après l'optimisation afin qu'ils puissent s'entraîner de nouveau sur les données train. Après avoir obtenu les prédictions, nous avons soumis nos fichiers de prédiction sur Kaggle pour connaître lequel de nos modèles nous fait avoir le meilleur score.



Graphique 14 - Classement Kaggle

---

Nous pouvons voir que le modèle qui se distingue positivement des autres est le modèle kernel ridge qui possède un score de 0,16816. A l'inverse, deux modèles se distinguent négativement, le modèle Gradient Boosting régressif et le modèle LightGBM régressif ayant un score inférieur à 0,16. Donc pour la suite, nous n'allons plus utiliser ces deux modèles.

### 13.4. Courbes d'apprentissage

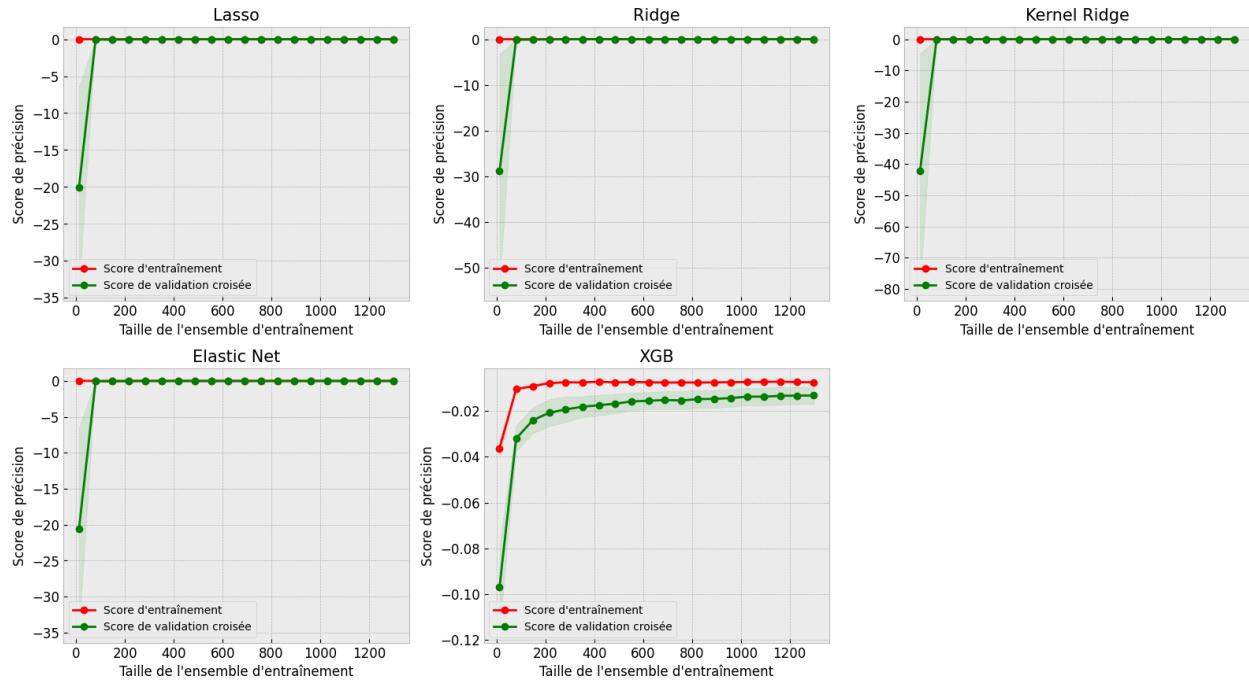
Pour affiner le nombre de modèles, nous allons utiliser les courbes d'apprentissage pour savoir quel modèle est le meilleur.

Noms des modèles				
Lasso	Ridge	Kernel Ridge	ElasticNet	XGBoost régressif

Tableau 11 - Liste des modèles restants

Les courbes de compromis biais-variance sont des outils essentiels en apprentissage automatique, permettant d'analyser le trade-off entre la capacité d'un modèle à capturer les caractéristiques des données (biais) et sa sensibilité aux fluctuations dans l'ensemble d'entraînement (variance). Leur utilité réside dans leur capacité à visualiser comment la complexité du modèle affecte ces deux aspects, aidant ainsi à trouver le point optimal où le compromis entre biais et variance est minimal, conduisant à une meilleure généralisation sur des données inconnues.

### Courbes d'apprentissage des modèles optimisés



Graphique 15 - Courbes d'apprentissage

Nous pouvons remarquer sur les graphiques précédents que le modèle XGBoost n'est pas aussi performant que les autres. En effet, les deux courbes score d'entraînement et score de validation croisée sont plus éloignées. A l'inverse, les quatres autres modèles, les courbes se superposent ce qui indique de très bonnes prédictions. Donc nous avons décidé de ne pas soumettre les prédictions du modèle XGBoost pour la compétition Kaggle.

**Remarque :** nous avons mis en [annexe 9](#), un graphique montrant l'importance de chaque variables dans certains modèles.

---

## 14. Choix du modèle final

L'apprentissage ensembliste, également connu sous le nom d'ensemble learning en anglais, est une technique en apprentissage automatique où plusieurs modèles d'apprentissage sont combinés pour améliorer les performances globales de prédiction. Plutôt que de s'appuyer sur un seul modèle, l'apprentissage ensembliste exploite la diversité des prédictions de plusieurs modèles pour obtenir des résultats plus robustes et précis. L'idée fondamentale derrière l'apprentissage ensembliste est que la combinaison de multiples modèles permet de compenser les faiblesses individuelles de chaque modèle, conduisant à une meilleure capacité de généralisation et à des performances globales plus élevées sur de nouveaux exemples.

❖ **Apprentissage ensembliste simple :**

Dans l'apprentissage ensembliste simple, les modèles individuels sont généralement entraînés de manière indépendante, et leurs prédictions sont ensuite combinées par des méthodes simples telles que la moyenne ou le vote majoritaire.

❖ **Apprentissage ensembliste avancé :**

L'apprentissage ensembliste avancé implique des techniques plus sophistiquées qui combinent les modèles individuels de manière plus complexe et adaptative pour maximiser les performances. Un exemple d'apprentissage ensembliste avancé comprend des techniques telle que le Stacking, qui implique l'utilisation d'un modèle de métaprédition qui prend en compte les prédictions de plusieurs modèles de base, en utilisant ces prédictions comme nouvelles caractéristiques pour prédire la cible finale.

Nous avons donc soumis les prédictions pour chaque modèle et c'est le modèle Kernel Ridge que nous avons retenu comme le meilleur.

---

## **15. Application de restitution visuelle des résultats**

### **❖ Présentation de l'outil**

Après avoir mené notre étude, nous avons développé un outil de restitution pour présenter les résultats de manière conviviale et intuitive. L'objectif principal de cet outil est d'aider les utilisateurs à estimer les prix des logements qu'ils envisagent d'acheter ou de vendre. Nous utiliserons les prédictions des prix des logements générées précédemment pour alimenter cet outil et fournir une assistance précieuse dans le processus de prise de décision.

L'outil propose à l'utilisateur trois perspectives d'analyse distinctes : une vue d'ensemble des statistiques du marché, une vue dédiée aux acheteurs et une autre aux vendeurs. La vue des statistiques globales offre une vision complète du marché immobilier, notamment en ce qui concerne les prix de vente des logements entre 2006 et 2010. Les utilisateurs peuvent consulter le prix moyen des transactions, ainsi que les moyennes de pièces et de chambres par logement. De plus, ils ont la possibilité de visualiser les prix moyens par quartier et en fonction de la superficie. Les logements sont également évalués selon leur qualité globale, avec des notes ou des avis disponibles pour guider les utilisateurs dans leur prise de décision.

Les perspectives axées sur les acheteurs et les vendeurs permettent aux utilisateurs de sélectionner les critères spécifiques du logement qu'ils envisagent d'acquérir ou de vendre. Ils peuvent obtenir des informations sur le nombre de logements disponibles, une fourchette de prix ainsi que la disponibilité d'un garage ou d'une piscine. En cas de vente, une estimation du prix de leur bien est fournie une fois qu'ils ont sélectionné les caractéristiques pertinentes.

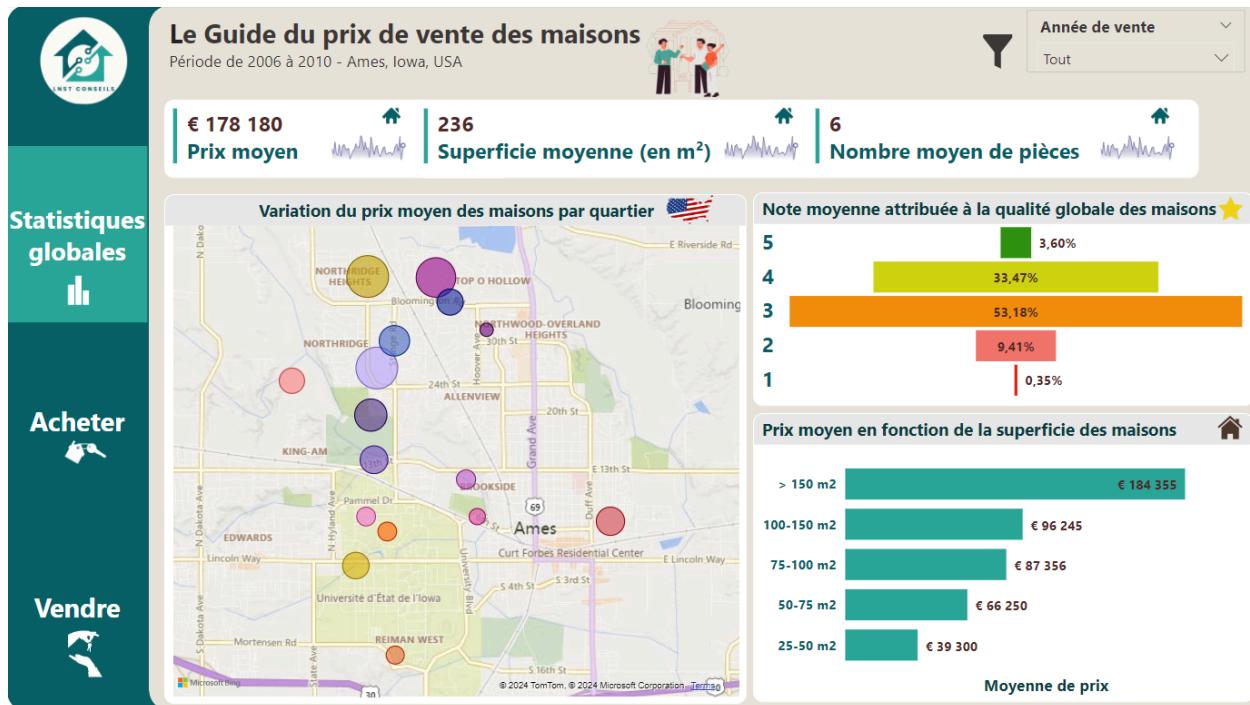


Figure 4 - Première page de l'outil



Figure 5 - Deuxième page de l'outil



Figure 6 - Troisième page de l'outil

#### ❖ Guide d'utilisation

L'outil créé sur Power BI est disponible en téléchargeant l'application Power BI Desktop. Son interface permet une visualisation des éléments sans nécessiter d'importation ou de rafraîchissement des données. Pour une expérience plus pratique et intuitive, un lien vers l'outil est également accessible sur ce site : <https://app.powerbi.com/groups/me/reports/446e217c-a788-44e6-9e7b-7cb18eb0a570/ReportSection6c45589fed89cae6c057?experience=power-bi>. Il vous suffit d'activer la licence gratuite de Power BI pour y accéder.

---

## Conclusion

La réalisation de ce rapport sur les prédictions des prix de vente des logements à Ames, Iowa, aux États-Unis, a nécessité un travail méthodique et rigoureux de la part de toute l'équipe impliquée. Nous avons pu mettre en place une structure solide, tant au niveau de l'organisation que des outils utilisés, notamment le Trello et le Git, qui ont facilité la collaboration et la gestion des tâches.

La présentation et la qualité des données ont été des étapes cruciales, où nous avons dû traiter les valeurs manquantes et aberrantes, ainsi que procéder à un travail minutieux de *feature engineering* pour enrichir notre base de données.

La construction et l'évaluation des différents modèles de *Machine Learning* ont été réalisées avec attention, en explorant une variété d'approches, de la régression linéaire aux méthodes plus complexes telles que les arbres de décision, la régression linéaire, le lasso, la machine à vecteurs de support, les k-plus proche voisins, le forêt aléatoire et les modèles de boosting. Chaque modèle a été interprété à travers des métriques telles que le coefficient de détermination et l'erreur quadratique moyenne de la racine, permettant ainsi de comparer leur performance.

Suite à cette évaluation approfondie, nous avons pu choisir le modèle final le plus performant pour prédire les prix des logements à Ames qui est celui de l'apprentissage ensembliste du modèle Kernel Ridge. Cette sélection a été étayée par une analyse approfondie des résultats et des critères tels que la précision et la robustesse du modèle. De plus, dans notre étude, il nous a permis d'obtenir le meilleur score sur Kaggle à savoir 0.16816 et une position de 679 / 4394.

Enfin, grâce à l'application de restitution visuelle des résultats Power BI, nous avons pu visualiser les données et définir de manière claire et intuitive les prédictions des prix de vente des logements, offrant ainsi une valeur ajoutée aux équipes métiers (pour plus d'informations, voir l'application Power BI).

---

En conclusion, ce rapport démontre l'efficacité d'une approche méthodique et rigoureuse dans la prédiction des prix de vente des logements aux États-Unis. Il met en lumière l'importance de la qualité des données, du choix et de l'évaluation des modèles, ainsi que de la restitution visuelle des résultats pour aboutir à des prédictions précises et exploitables. Ce travail représente une contribution significative dans le domaine de l'analyse immobilière et offre des perspectives prometteuses pour de futures applications et recherches.

## Annexes

### Annexe 1 : Dictionnaire des variables

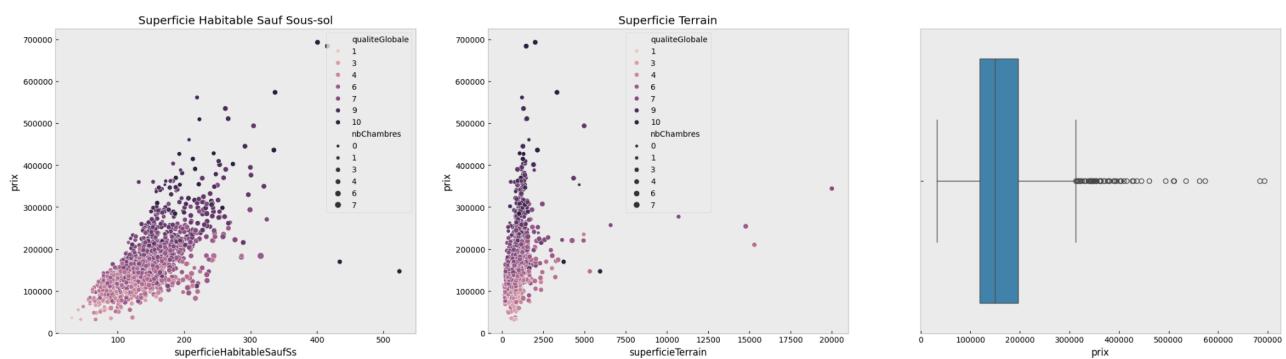
Variables	Renommage	Type	Description
1stFlrSF	superficieEtage1	float	Pieds carrés du premier étage
2ndFlrSF	superficieEtage2	float	Pieds carrés au deuxième étage
3SsnPorch	superficieRallonge3Saisons	float	Superficie du porche trois saisons en pieds carrés
Alley	typeAlleeAcces	string	Type de ruelle d'accès à la propriété
BedroomAbvGr	nbChambres	integer	Chambres au-dessus du sol (n'inclut PAS les chambres du sous-sol)
BldgType	typeBatiment	string	Type de logement
BsmtCond	conditionSousSol	string	Évalue l'état général du sous-sol
BsmtExposure	expositionSousSol	string	Fait référence aux murs de sortie ou de rez-de-jardin
BsmtFinSF1	superficieFinieSousSol1	float	Pieds carrés finis de type 1
BsmtFinSF2	superficieFinieSousSol2	float	Pieds carrés finis de type 2
BsmtFinType1	qualiteSurfaceFinieSousSol1	string	Evaluation de la surface finie du sous-sol
BsmtFinType2	qualiteSurfaceFinieSousSol2	string	Evaluation de la surface finie du sous-sol (si plusieurs types)
BsmtFullBath	salleBainSs	integer	salles de bain complètes au sous-sol
BsmtHalfBath	demiSalleBainSs	integer	demi-salles de bain au sous-sol
BsmtQual	qualiteSousSol	string	Évalue la hauteur du sous-sol
BsmtUnfSF	superficieSousSolNonAmenagee	float	Pieds carrés non aménagés de sous-sol
CentralAir	climatisation	string	Climatisation centrale
Condition1	proxRoute1	string	Proximité de diverses conditions
Condition2	proxRoute2	string	Proximité de diverses conditions (si plusieurs conditions sont présentes)
Electrical	systElectrique	string	Système électrique
EnclosedPorch	superficiePorcheFerme	float	superficie du porche fermé en pieds carrés

ExterCond	conditionExterieur	string	Evalue l'état actuel du matériau à l'extérieur
Exterior1st	materiauExterieur1	string	Revêtement extérieur sur maison
Exterior2nd	materiauExterieur2	string	Revêtement extérieur de la maison (si plus d'un matériau)
ExterQual	qualiteExterieur	string	Évalue la qualité du matériau à l'extérieur
Fence	cloture	string	Qualité de clôture
FireplaceQu	qualiteCheminée	string	Qualité de cheminée
Fireplaces	nbCheminées	integer	Nombre de cheminées
Foundation	fondation	string	Type de fondation
FullBath	nbSalleBain	integer	salles de bains complètes au-dessus du niveau du sol
Functional	fonctionnalites	string	fonctionnalité domestique (supposer typique, sauf si des déductions sont justifiées)
GarageArea	superficieGarage	float	Taille du garage en pieds carrés
GarageCars	nbPlacesVoiture	integer	Taille du garage en capacité de voiture
GarageCond	conditionGarage	string	État du garage
GarageFinish	interieurGarage	string	Finition intérieure du garage
GarageQual	qualiteGarage	string	la qualité du garage
GarageType	typeGarage	string	Emplacement du garage
GarageYrBlt	anneeConstrGarage	integer	Année de construction du garage
GrLivArea	superficieHabitableSaufSs	float	Surface habitable hors sous sol en pieds carrés
HalfBath	nbDemiSalleBain	integer	demi-bains au-dessus du niveau du sol
Heating	chauffage	string	Type de chauffage
HeatingQC	qualiteChauffage	string	Qualité et état du chauffage
HouseStyle	styleBatiment	string	Style d'habitation
Id	id	integer	identifiant
KitchenAbvGr	nbCuisines	integer	Cuisines au-dessus du niveau du sol
KitchenQual	qualiteCuisine	string	la qualité de la cuisine

LandContour	contourTerrain	string	Planéité de la propriété
LandSlope	penteTerrain	string	Pente de la propriété
LotArea	superficieTerrain	float	Taille du terrain en pieds carrés
LotConfig	configTerrain	string	Configuration des lots
LotFrontage	longTerrainRue	float	Pieds linéaires de rue reliés à la propriété
LotShape	formeTerrain	string	Forme générale de la propriété
LowQualFinSF	superficieQualiteInferieure	float	Pieds carrés finis de faible qualité (tous les étages)
MasVnrArea	superficiePlacageMaconnerie	float	superficie de placage de maçonnerie en pieds carrés
MasVnrType	typePlacageMaconnerie	string	Type de placage de maçonnerie
MiscFeature	elementsDivers	string	fonctionnalité diverse non couverte dans les autres catégories
MiscVal	valeursElementsDivers	float	\$Valeur de la fonctionnalité diverse
MoSold	moisVente	integer	Mois de vente (MM)
MSZoning	zonage	string	Identifie la classification générale de zonage de la vente.
Neighborhood	quartier	string	Emplacements physiques dans les limites de la ville d'Ames
OpenPorchSF	superficiePorcheOuvert	float	superficie du porche ouvert en pieds carrés
OverallCond	contitionGlobale	string	Evalue l'état général de la maison
OverallQual	qualiteGlobale	string	Evalue l'ensemble des matériaux et de la finition de la maison
PavedDrive	alleePavee	string	allée pavée
PoolArea	superficiePiscine	float	Superficie de la piscine en pieds carrés
PoolQC	qualitePiscine	string	Qualité de la piscine
RoofMatl	materiauToit	string	Matériau de toiture
RoofStyle	styleToit	string	Type de toit
SaleCondition	conditionVente	string	Condition de vente
SalePrice	prix	float	Prix du logement
SaleType	typeVente	string	Type de vente

ScreenPorch	superficieSolarium	float	superficie du porche grillagé en pieds carrés
Street	typeRouteAcces	string	Type de route d'accès à la propriété
MSSubClass	typeClasseBatiment	string	Identifie le type de logement concerné par la vente.
TotalBsmtSF	superficieTotaleSousSol	float	Total en pieds carrés de superficie du sous-sol
TotRmsAbvGrd	nbPieces	integer	nombre total de chambres au-dessus du niveau du sol (n'inclut pas les salles de bains)
Utilities	utilites	string	Type d'utilitaires disponibles
WoodDeckSF	superficieTerrasseBois	float	Superficie de la terrasse en bois en pieds carrés
YearBuilt	anneeConstruction	integer	Date de construction originale
YearRemodAdd	anneeRenovation	integer	Date de rénovation (identique à la date de construction s'il n'y a pas de rénovation ou d'ajouts)
YrSold	anneeVente	integer	Année de vente (AAAA)

## Annexe 2 : Représentation graphique des valeurs aberrantes

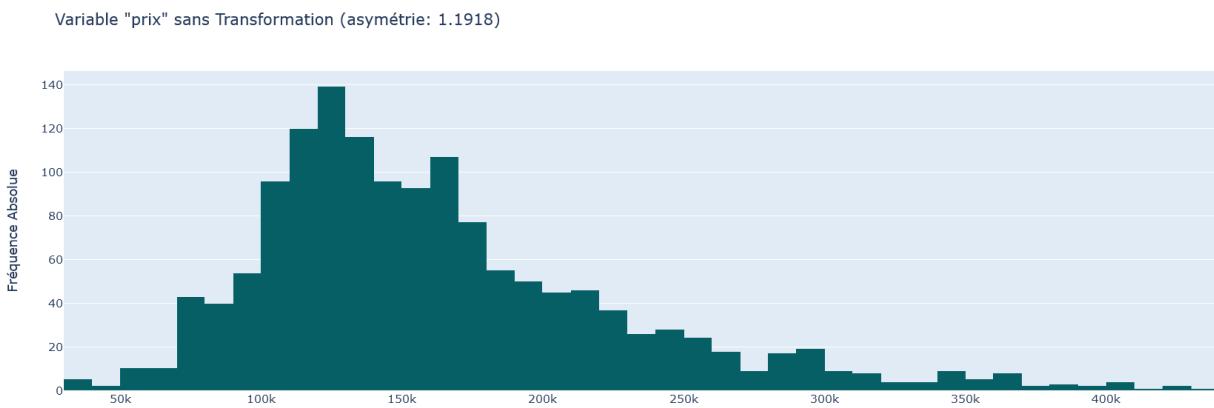


---

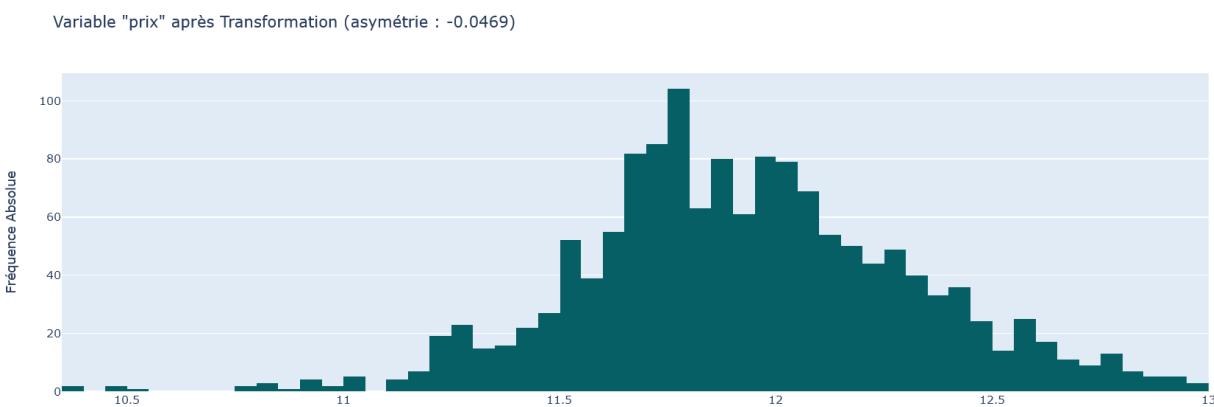
## Annexe 3 : Asymétrie

Une asymétrie dans les données correspond à une distribution asymétrique des données. En d'autre terme, la distribution des données n'est pas symétrique autour de sa moyenne.

### ❖ Distribution du prix avec la présence d'une asymétrie :



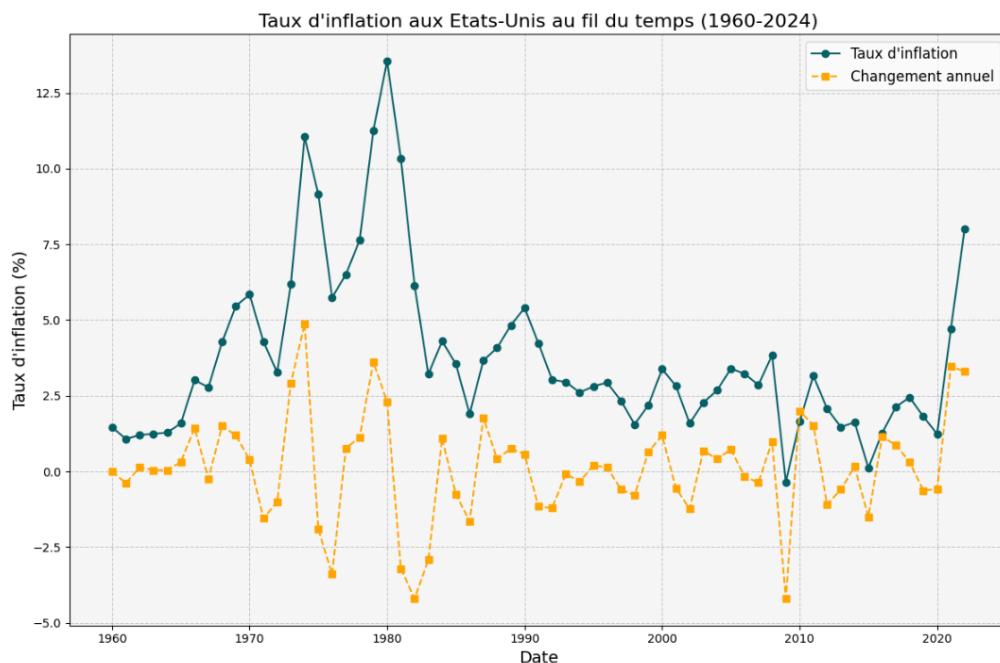
### ❖ Distribution du prix avec une réduction de l'asymétrie :



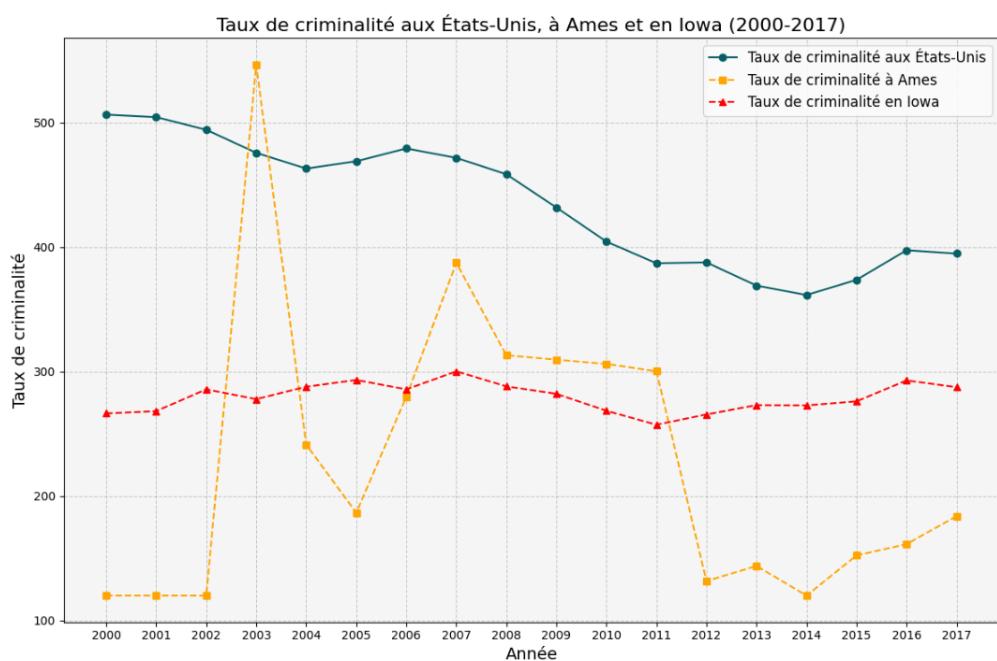
Une asymétrie de -0.0469 indique une distribution légèrement asymétrique mais pratiquement symétrique. Cela signifie que la distribution des données est presque équilibrée, avec peu ou pas de prédominance pour les valeurs élevées ou basses. En d'autres termes, il n'y a pas de tendance marquée à avoir plus de valeurs élevées ou basses dans les données.

**Remarque :** nous avons réalisé cette transformation pour toutes les variables numériques que nous avons dans notre base de données.

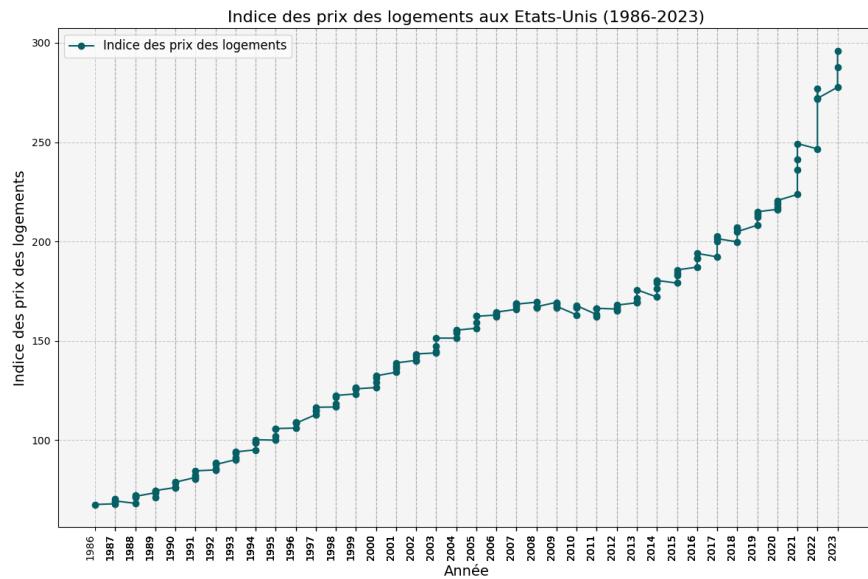
#### Annexe 4 : Taux d'inflation aux Etats-Unis



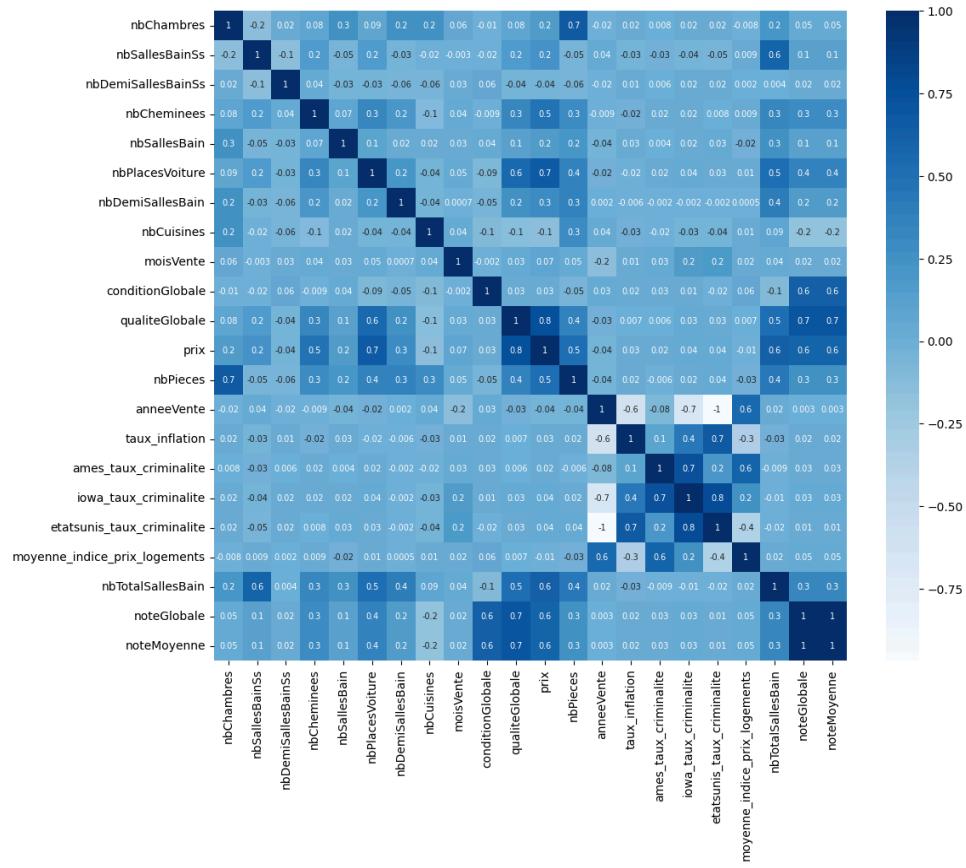
#### Annexe 5 : Taux de criminalité aux Etats-Unis



## Annexe 6 : Indice des prix des logements aux Etats-Unis



## Annexe 7 : Matrice de corrélation entre les variables explicatives numériques et la variable cible



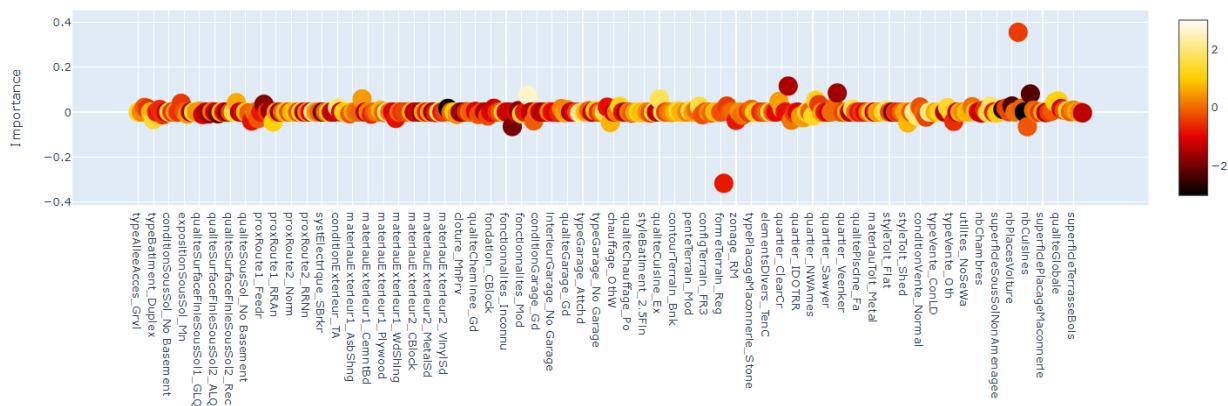
## Annexe 8 : Résultat ANOVA

Variables	Statistique ANOVA F Valeur	P	Signification
typeClasseBatiment	114.359006	7.998741e-47	Rejeter
zonage	80.036903	1.094003e-33	Rejeter
typeRouteAcces	4.483368	3.439700e-02	Rejeter
typeAlleeAcces	16.636005	7.198180e-08	Rejeter
formeTerrain	108.968389	1.214420e-24	Rejeter
contourTerrain	0.821419	4.400129e-01	Ne pas rejeter
utilites	0.318766	5.724378e-01	Ne pas rejeter
configTerrain	7.210409	8.376341e-05	Rejeter
penteTerrain	1.388486	2.497861e-01	Ne pas rejeter
quartier	73.785270	4.904471e-112	Rejeter
proxRoute1	16.535410	1.447444e-10	Rejeter
proxRoute2	6.705229	1.710917e-04	Rejeter
typeBatiment	21.962450	4.020809e-10	Rejeter
styleBatiment	32.175609	1.026129e-25	Rejeter
styleToit	29.498290	2.782015e-13	Rejeter
materiauToit	10.635103	1.135531e-03	Rejeter
materiauExterieur1	45.977604	3.697481e-44	Rejeter
materiauExterieur2	43.302640	1.118262e-41	Rejeter
typePlacageMaconnerie	155.239306	8.794522e-62	Rejeter
qualiteExterieur	460.650556	8.952321e-210	Rejeter
conditionExterieur	20.031322	2.625689e-09	Rejeter
fondation	183.553022	8.837148e-101	Rejeter
conditionSousSol	38.732127	4.124806e-17	Rejeter
expositionSousSol	55.575448	1.259026e-43	Rejeter
qualiteSurfaceFinieSousSol1	70.957705	1.455337e-77	Rejeter
qualiteSurfaceFinieSousSol2	8.241430	8.623404e-09	Rejeter
chauffage	12.626504	3.663313e-06	Rejeter
qualiteChauffage	97.899009	9.539246e-74	Rejeter
climatisation	113.303939	1.594270e-25	Rejeter
systElectrique	102.823911	2.182029e-23	Rejeter
qualiteCuisine	218.914804	9.583623e-84	Rejeter

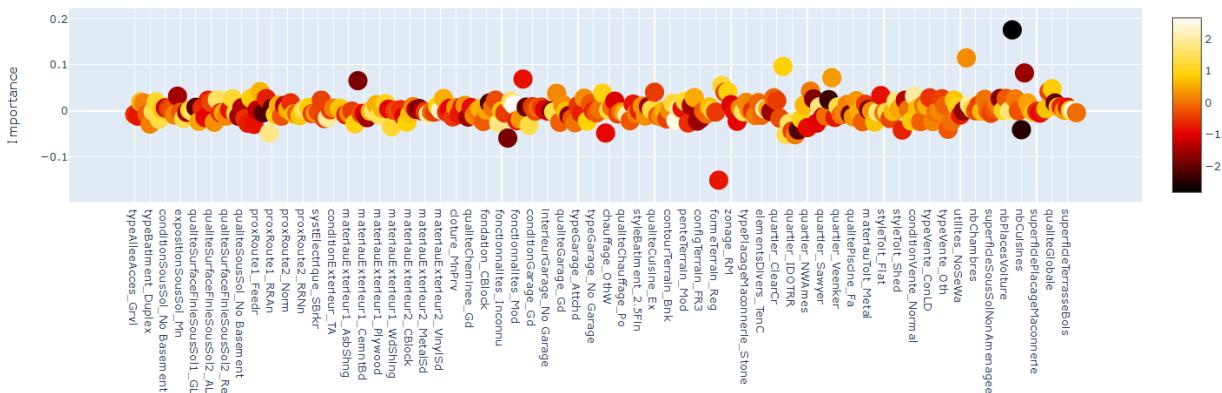
fonctionnalites	6.533068	3.290621e-05	Rejeter
qualiteCheminée	129.240480	3.255371e-113	Rejeter
typeGarage	132.074324	3.745938e-96	Rejeter
interieurGarage	234.285574	8.827508e-124	Rejeter
qualiteGarage	29.924128	5.301002e-29	Rejeter
conditionGarage	137.499615	2.141745e-30	Rejeter
alleePavee	47.418689	1.136731e-20	Rejeter
qualitePiscine	0.672397	5.689933e-01	Ne pas rejeter
cloture	16.047058	7.328363e-13	Rejeter
elementsDivers	3.264022	1.124536e-02	Rejeter
typeVente	73.994393	1.644220e-44	Rejeter
conditionVente	58.967537	3.804497e-46	Rejeter
route_ville	1.435755	2.310244e-01	Ne pas rejeter
autoroute	44.223525	4.145571e-11	Rejeter
proximite_gare	0.267153	6.053270e-01	Ne pas rejeter
proximite_parc	11.441212	7.375658e-04	Rejeter
hauteurSs	313.864806	3.365651e-194	Rejeter

## Annexe 9 : Importance des caractéristiques

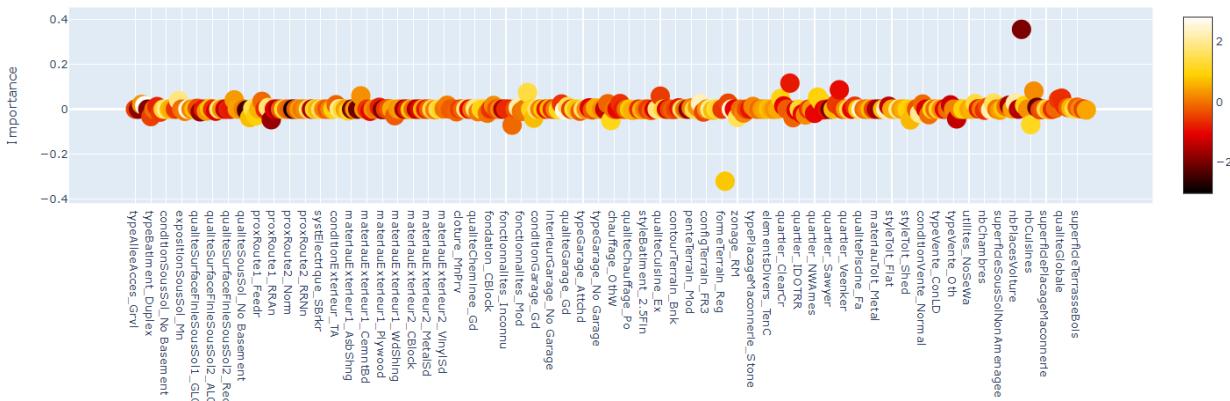
Importance des caractéristiques de Lasso



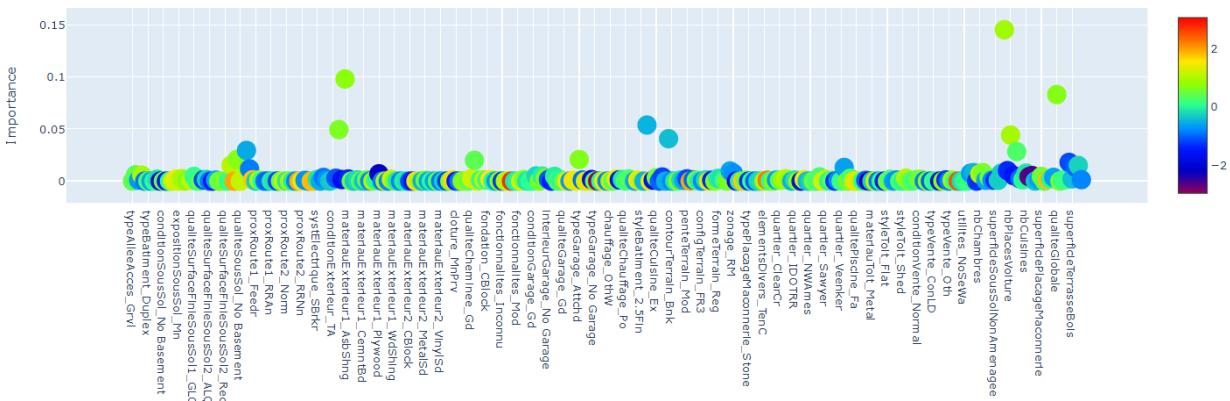
### Importance des caractéristiques de Ridge



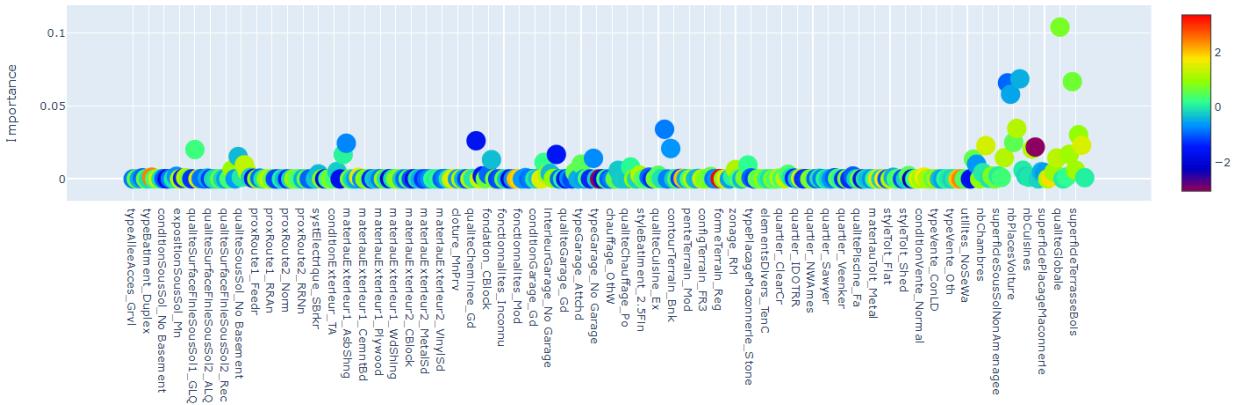
### Importance des caractéristiques de ElasticNet



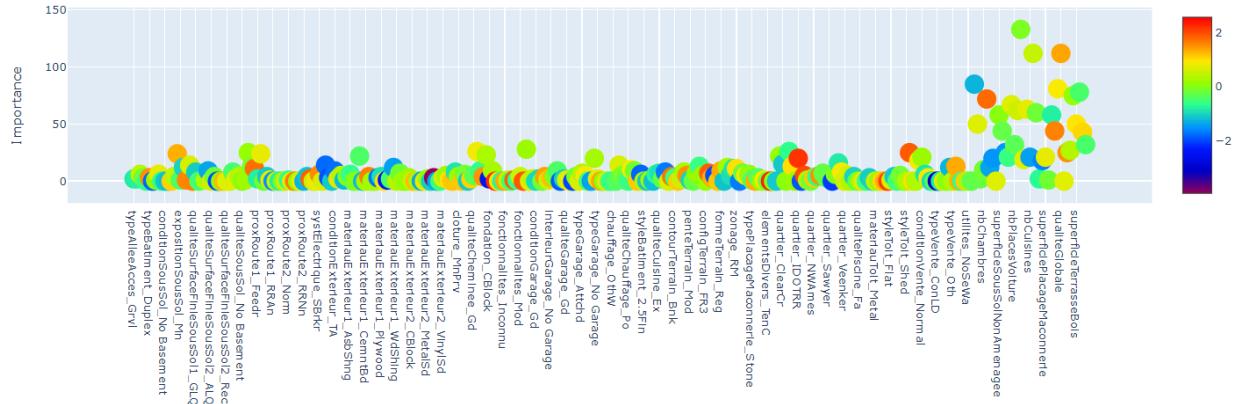
### Importance des caractéristiques de XGB



### Importance des caractéristiques de GB

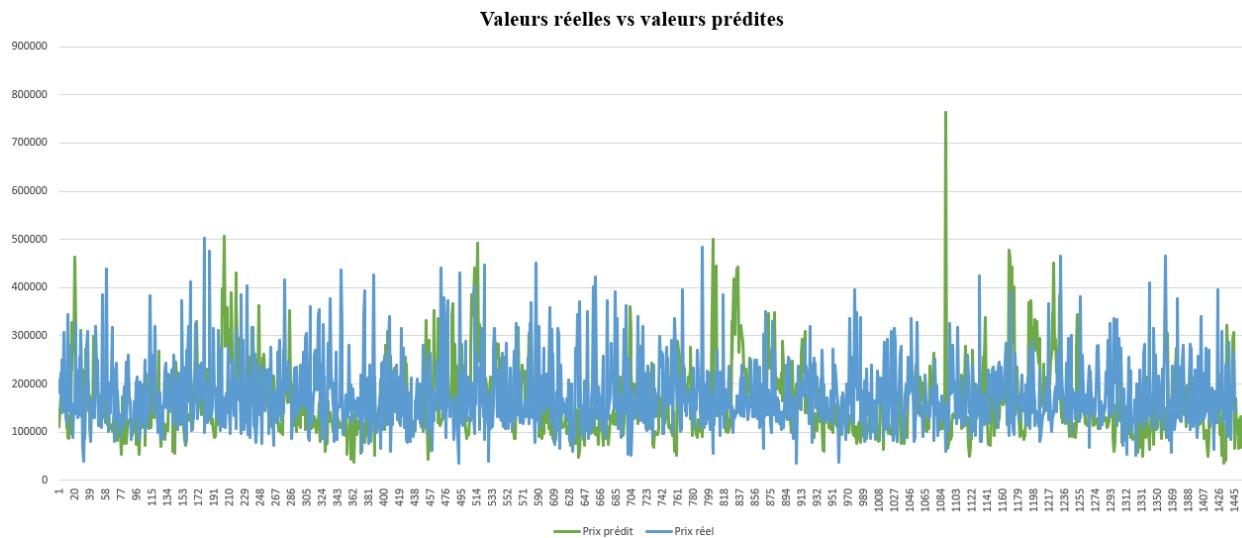


### Importance des caractéristiques de LGB



---

## Annexe 10 : Résultat de la prédiction



---

## Rapport métier

### Quelques chiffres clés

Nous observons que la plupart des logements ont des prix compris entre 100 000€ et 200 000€. La répartition des prix des logements est assez diversifiée, avec une tendance à se concentrer autour des prix moyens.

En outre, les prix des logements varient de 32610 \$ à 468 913\$, avec une moyenne de 183 381\$. Le fait que la moyenne des prix des logements ne soit pas très élevée malgré la large gamme de prix observée suggère une certaine diversité et accessibilité sur le marché immobilier, ce qui peut être perçu comme positif pour les acheteurs potentiels (voir figure 1).



Figure 4 - Représentation visuelle de quelques résumés numériques

### Analyse de la distribution de quelques paramètres clés

Penchons-nous maintenant sur une analyse détaillée de la répartition de certains paramètres clés essentiels pour une recherche ou une vente de logements, et qui peuvent également avoir un impact sur le prix de vente d'un bien immobilier.

## ❖ Distribution des logements en fonction de la superficie habitable :

La majorité des logements (environ 69 %) ont une superficie habitable entre 100 et 200 mètres carrés, avec près de la moitié (42.5%) située entre 100 et 150 mètres carrés, et environ un quart (27%) entre 150 et 200 mètres carrés. Les logements très spacieux (plus de 250 mètres carrés) et très petits (moins de 50 mètres carrés) sont rares, représentant respectivement environ 2 % et 0,5 % de l'échantillon (voir figure 2).

Cette répartition suggère que la majorité des logements de l'Iowa se situent dans une gamme de taille moyenne à grande, tandis que les logements très petits ou très grands sont moins fréquents.

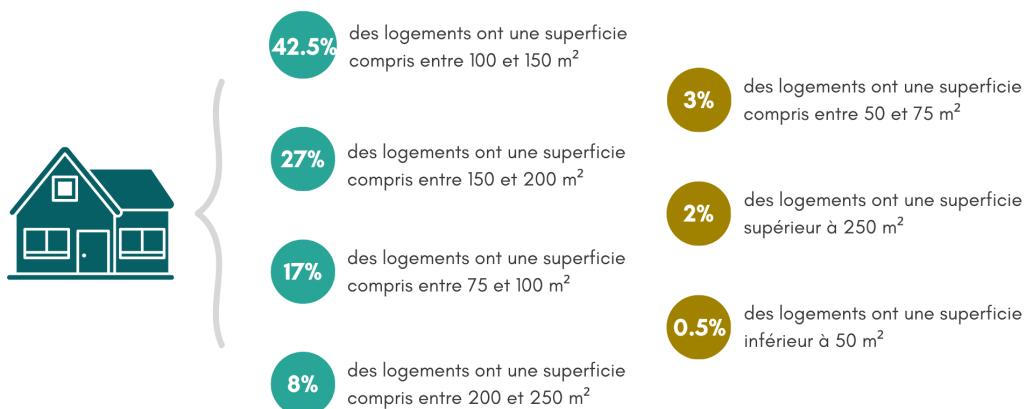


Figure 5 - Représentation visuelle de la distribution

## ❖ Distribution des logements en fonction du nombre de pièces :

La majorité des logements ont entre 5 et 7 pièces, avec plus de la moitié des logements ayant 6 ou 7 pièces. Les logements avec 8 pièces sont également assez présents, représentant environ 13% du total. Les logements avec moins de 3 pièces ou plus de 10 pièces sont moins courants, chacun représentant moins de 3% du total (voir figure 3).

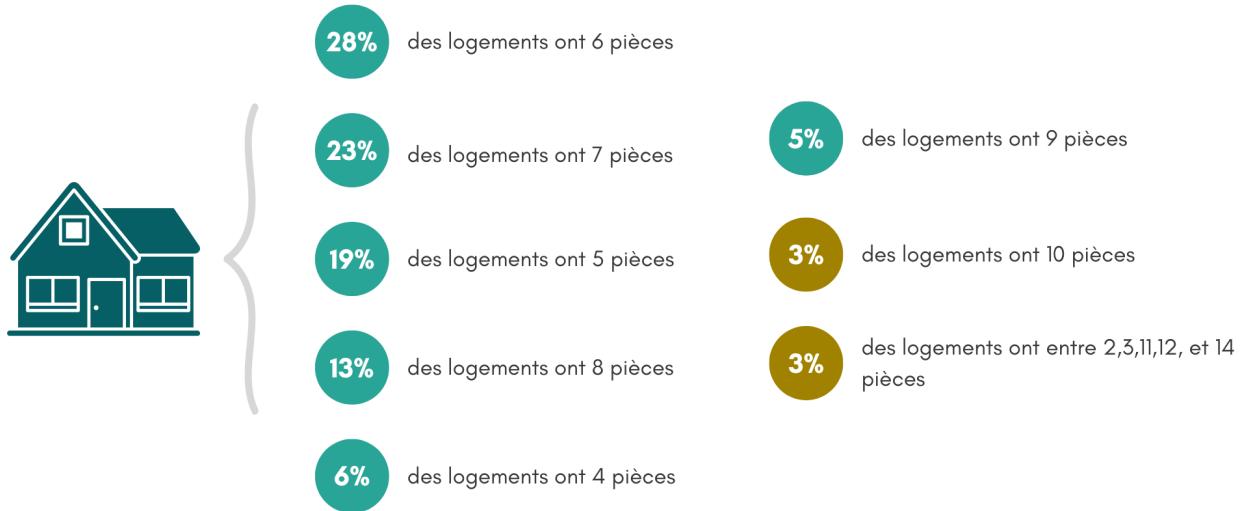


Figure 6 - Représentation visuelle de la distribution

#### ❖ Distribution des logements en fonction de la qualité globale :

L'analyse du marché immobilier révèle une prédominance de maisons de qualité modeste à moyenne. Les logements sont notés sur une échelle de 1 à 10, et aucun n'a obtenu une note supérieure à 5, indiquant ainsi l'absence de ventes de maisons haut de gamme à Iowa durant la période étudiée. Cette situation s'explique par le fait que les logements les plus luxueux sont généralement hors de portée pour une grande partie de la population (voir figure 4).

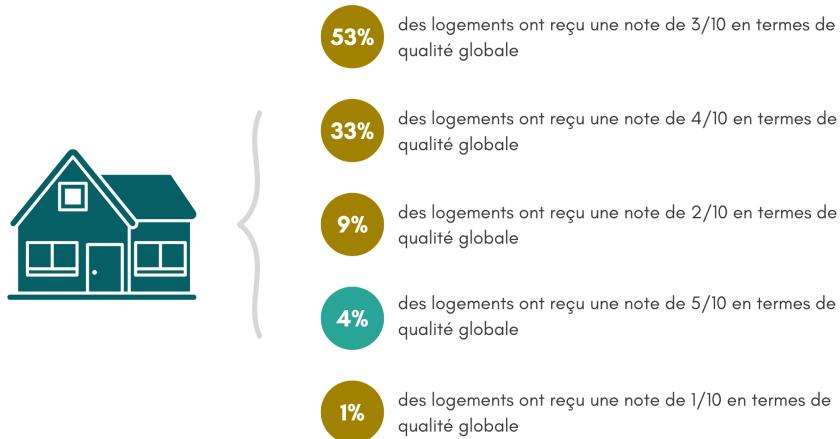


Figure 7 - Représentation visuelle de la distribution

Dans le but de comprendre comment évolue le prix des logements en fonction de différents paramètres, nous allons étudier sa répartition en fonction de ces derniers.

#### ❖ Distribution du prix en fonction de l'année de vente :

En analysant les prix moyens des ventes de maisons sur la période de 2006 à 2010, il est notable qu'une tendance à la hausse était présente de 2006 à 2007, avec un pic atteint en 2007 à 170 000.72 €. Cependant, cette tendance a été interrompue par la crise financière de 2007 à 2008, où l'on observe une diminution significative du prix moyen en 2008 (165 618.96 €). Cette baisse se poursuit en 2009 avec un prix moyen de 164 941.75 €, reflétant l'impact négatif de la crise sur le marché immobilier (voir figure 5).

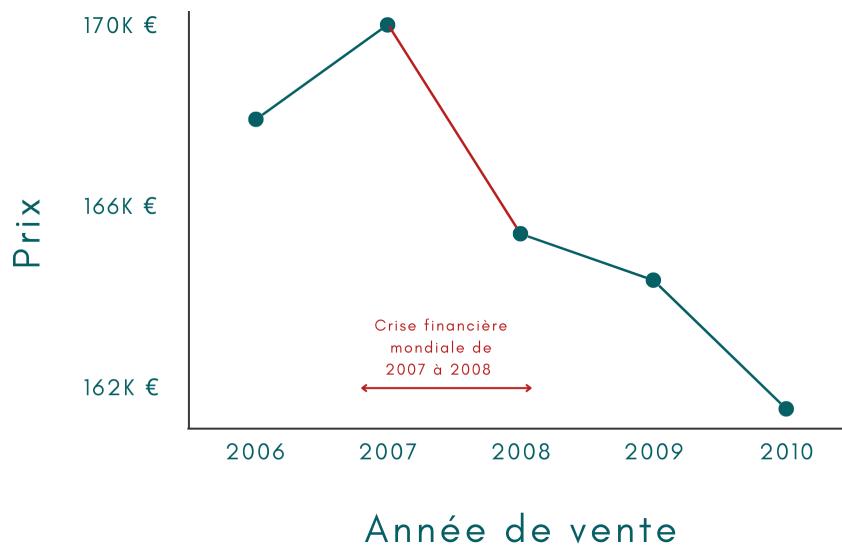


Figure 8 - Représentation visuelle de la distribution

#### ❖ Distribution du prix en fonction de la superficie totale :

En analysant la relation entre le prix et la superficie totale des maisons, nous observons une corrélation positive entre ces deux variables. En d'autres termes, à mesure que la superficie totale des logements augmente, le prix tend à augmenter également (voir figure 6).

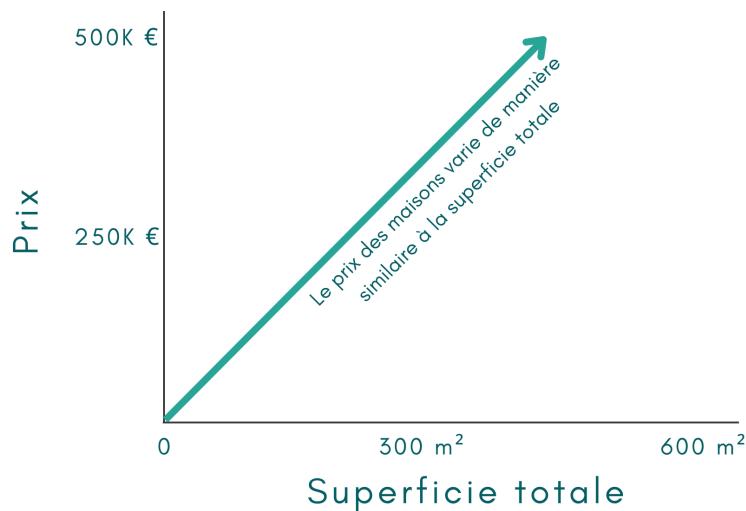


Figure 9 - Représentation Visuelle

---

#### **❖ Distribution du prix en fonction de l'ancienneté de rénovation :**

Contrairement à la superficie totale, le coût des logements diminue à mesure que leur date de rénovation recule dans le temps. Cependant, il est important de noter que cette diminution de prix est plus marquée pour les logements rénovés depuis plus de 30 ans. Par conséquent, les logements neufs sont les plus coûteux.

#### **❖ Distribution du prix en fonction du nombre de pièces des logements :**

L'analyse de la relation entre le prix et le nombre de pièces des maisons révèle une tendance où le prix augmente au fur et à mesure que le nombre de pièces augmente, suggérant ainsi une relation positive entre ces deux paramètres. Toutefois, cette augmentation de prix n'est pas uniforme. Elle est modérée pour les logements comptant moins de 3 pièces, puis s'accélère entre 4 et 8 pièces avant de se stabiliser à nouveau. Ceci évoque que la différence de prix entre les maisons ayant moins de 2 pièces n'est pas significative, tandis que cette différence devient plus notable entre 4 et 8 pièces.

Toutes ces informations concernant les caractéristiques pertinentes des logements et le prix du marché sont indispensables tant pour les acheteurs que pour les vendeurs immobiliers. Elles permettent aux acheteurs de mieux appréhender le prix moyen des maisons, les éléments qui influent sur les prix, et ainsi de négocier de manière plus éclairée. Quant aux vendeurs, ces données leur offrent une vision claire des tendances du marché, des critères à considérer pour établir un prix de vente, ainsi que des stratégies pour attirer efficacement les acheteurs en mettant en avant les caractéristiques les plus recherchées.

---

## Analyse des données externes

Nous allons désormais analyser la répartition des données externes que nous avons ajouté à l'étude, afin d'évaluer leur influence sur les prix des maisons.

### ❖ Distribution du prix moyen et du taux d'inflation en fonction de l'année de vente :

Nous sommes conscients que le taux d'inflation peut influencer les prix des logements sur le marché. En effet, une inflation élevée peut se traduire par une augmentation des coûts de construction, y compris ceux des matériaux et de la main-d'œuvre. Elle peut également réduire le pouvoir d'achat des ménages et entraîner une tendance à la hausse des prix des logements. C'est pourquoi nous souhaitons voir comment le prix moyen des maisons évolue en fonction de l'inflation par année.

Année de vente	Prix moyen	Taux d'inflation
2006	168128	3,2
2007	170011	2,9
2008	165619	3,8
2009	164942	-0,4
2010	161462	1,6

Tableau 10 : Distribution du prix moyen et du taux d'inflation en fonction de l'année de vente

Nous observons une variation du prix moyen des maisons d'une année à l'autre, mais cette variation ne semble pas être directement liée au taux d'inflation. Par exemple, en 2007, bien que le taux d'inflation ait diminué par rapport à l'année précédente, le prix moyen des maisons a

---

légèrement augmenté. De même, en 2009, malgré un taux d'inflation négatif, le prix moyen des maisons est resté relativement stable.

Ces observations suggèrent qu'il n'existe pas de relation claire et directe entre le prix des logements et le taux d'inflation dans notre ensemble de données. Cependant, d'autres facteurs peuvent également influencer les prix des logements. Nous prévoyons donc d'utiliser des techniques de régression avancées pour évaluer plus précisément ce lien.

Il convient aussi de noter que la période d'observation est relativement courte (5 ans), ce qui pourrait limiter la portée de nos conclusions. Pour une analyse plus approfondie, il serait nécessaire d'étudier cette relation sur une période plus longue et de tenir compte d'autres facteurs macroéconomiques. De plus, la crise mondiale entre 2007 et 2008 a probablement eu un impact significatif sur le marché immobilier, ce qui pourrait également influencer les résultats de notre analyse.

#### **❖ Distribution du prix et du taux de criminalité dans la ville d'Ames en fonction du quartier :**

Le taux de criminalité peut avoir un impact sur le prix des logements sur le marché. Les acheteurs potentiels sont généralement sensibles à la sécurité et à la qualité de vie dans un quartier donné. Ainsi, un taux de criminalité élevé peut dissuader les acheteurs et entraîner une diminution de la demande de logements dans cette zone. De plus, une augmentation du taux de criminalité peut influencer les perceptions des acheteurs concernant la sécurité de leur investissement immobilier, ce qui peut conduire à une dépréciation de la valeur des logements dans les quartiers touchés. En conséquence, les vendeurs peuvent être contraints de réduire les prix pour attirer des acheteurs ou pour compenser les risques perçus associés à la criminalité dans la région.

Après avoir étudié la relation entre le prix moyen des logements et le taux d'inflation par quartier, nous avons observé que, de manière similaire à ce que nous avons constaté pour le taux

---

d'inflation, les données disponibles ne révèlent pas de relation linéaire claire entre le taux de criminalité moyen et le prix moyen des logements dans les quartiers répertoriés. Par exemple, bien que le quartier *Somerset* affiche le taux de criminalité moyen le plus élevé, il présente également l'une des moyennes de prix les plus élevées, tandis que le quartier North Ames présente un taux de criminalité moyen similaire mais des prix moyens nettement plus bas. Ces observations suggèrent qu'il existe d'autres facteurs influençant les prix des logements dans ces quartiers, en plus du taux de criminalité. Une étude plus approfondie utilisant des techniques statistiques, que nous entreprendrons par la suite, nous permettra d'explorer davantage cette relation.

## Résultat de la prédiction des prix des logements

La modélisation utilisant des méthodes de régression avancées pour prédire le prix des logements à Ames, Iowa, consiste à appliquer des techniques statistiques sophistiquées pour analyser les relations entre les différentes variables des maisons (telles que la taille, l'emplacement, les caractéristiques structurelles, etc.) et leur prix de vente. Ces méthodes avancées de régression vont au-delà des approches traditionnelles et simples de régression linéaire, en utilisant par exemple des techniques telles que la régression ridge, la régression lasso, les arbres de décision, et les forêts aléatoires.

L'objectif est de développer un modèle qui puisse capturer au mieux la complexité des données immobilières et fournir des prédictions précises du prix des logements à partir des caractéristiques disponibles. Ces modèles avancés peuvent être calibrés et validés à l'aide de techniques telles que la validation croisée et l'optimisation des hyperparamètres pour assurer leur performance et leur généralisabilité.

Ainsi, après avoir mené une analyse des données immobilières de la ville d'Ames, Iowa, et développé plusieurs modèles de prédiction, nous avons obtenu des résultats prometteurs. Les modèles de prédiction ont été évalués en utilisant des métriques telles que le coefficient de

---

détermination ( $R^2$ ) et l'erreur quadratique moyenne de la racine (RMSE) pour mesurer leur précision et leur performance.

Nous avons constaté que les modèles de régression linéaire, Lasso, Ridge, ainsi que les méthodes basées sur les arbres de décision et les forêts aléatoires ont tous été capables de prédire les prix de ventes des logements avec une précision raisonnable.

Après une analyse approfondie des résultats, l'apprentissage ensembliste du modèle Kernel Ridge a été choisi comme le modèle final pour la prédiction des prix de ventes des logements à Ames, Iowa. Ce choix a été motivé par sa capacité à capturer la complexité des données et à fournir des prédictions précises et fiables mais également par sa capacité à fournir le meilleur score possible par rapport aux autres.

Après avoir effectué cette modélisation, les valeurs prédictives sont très similaires à celles observées, avec une marge d'erreur minime, démontrant ainsi la robustesse de notre étude ([Annexe 10 : Résultat de la prédiction](#)).

Les résultats de notre analyse offrent des perspectives prometteuses pour les professionnels de l'immobilier, les investisseurs et les décideurs, en fournissant des outils précieux pour évaluer et prédire les prix des logements à Ames, Iowa. Ces informations peuvent également être utilisées pour prendre des décisions éclairées en matière d'achat, de vente ou d'investissement immobilier dans la région.

---

## Bibliographie :

- ❖ <https://medium.com/analytics-vidhya/advanced-regression-techniques-to-predict-home-prices-with-python-86caa9e0861d>
  - ❖ <https://www.kaggle.com/code/eraaz1/a-comprehensive-guide-to-advanced-regression>
  - ❖ [https://shire.science.uq.edu.au/CONS7008/\\_book/linear-models---analysis-of-variance-and-anova.html](https://shire.science.uq.edu.au/CONS7008/_book/linear-models---analysis-of-variance-and-anova.html)
  - ❖ <https://www.kaggle.com/code/carlmcbrideellis/house-prices-how-to-work-offline>
  - ❖ [https://shire.science.uq.edu.au/CONS7008/\\_book/linear-models---analysis-of-variance-and-anova.html](https://shire.science.uq.edu.au/CONS7008/_book/linear-models---analysis-of-variance-and-anova.html)
  - ❖ <https://www.mygreatlearning.com/blog/understanding-of-lasso-regression/#:~:text=LASSO%20regression%2C%20also%20known%20as,Absolute%20Shrinkage%20and%20Selection%20Operator.>
  - ❖ <https://www.ibm.com/topics/lasso-regression>
  - ❖ <https://www.ibm.com/topics/ridge-regression#:~:text=Ridge%20regression%20is%20a%20statistical,regularization%20for%20linear%20regression%20models.>
  - ❖ [https://scikit-learn.org/stable/modules/generated/sklearn.linear\\_model.Ridge.html](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html)
  - ❖ [https://scikit-learn.org/stable/modules/generated/sklearn.kernel\\_ridge.html](https://scikit-learn.org/stable/modules/generated/sklearn.kernel_ridge.html)
  - ❖ <https://mlweb.loria.fr/book/en/kernelridgeregression.html>
  - ❖ <https://medium.com/@shruti.dhumne/elastic-net-regression-detailed-guide-99dce30b8e6e>
  - ❖ [https://scikit-learn.org/stable/modules/generated/sklearn.linear\\_model.ElasticNet.html](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.ElasticNet.html)
  - ❖ <https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html>
  - ❖ <https://www.geeksforgeeks.org/python-decision-tree-regression-using-sklearn/>
  - ❖ <https://www.analyticsvidhya.com/blog/2020/03/support-vector-regression-tutorial-for-machine-learning/>
  - ❖ <https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsRegressor.html>
  - ❖ <https://towardsdatascience.com/random-forest-regression-5f605132d19d>
-

- 
- ❖ <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostRegressor.html>
  - ❖ <https://towardsdatascience.com/all-you-need-to-know-about-gradient-boosting-algorithm-part-1-regression-2520a34a502>
  - ❖ <https://machinelearningmastery.com/xgboost-for-regression/>
  - ❖ <https://www.geeksforgeeks.org/regression-using-lightgbm/>
  - ❖ [https://medium.com/@sarita\\_68521/understanding-the-bias-variance-tradeoff-in-machine-learning-examples-and-solutions-5de459ddeabd#:~:text=The%20Tradeoff%3A%20Balancing%20Bias%20and%20Variance&text=This%20tradeoff%20is%20often%20visualized,error%20on%20a%20validation%20dataset.](https://medium.com/@sarita_68521/understanding-the-bias-variance-tradeoff-in-machine-learning-examples-and-solutions-5de459ddeabd#:~:text=The%20Tradeoff%3A%20Balancing%20Bias%20and%20Variance&text=This%20tradeoff%20is%20often%20visualized,error%20on%20a%20validation%20dataset.)
  - ❖ <https://www.analyticsvidhya.com/blog/2018/06/comprehensive-guide-for-ensemble-models#:~:text=Ensemble%20learning%20is%20a%20machine,collective%20intelligence%20of%20the%20ensemble.>
  - ❖ Practical Statistics for Data Scientists, 2nd Edition, Peter Bruce, Andrew Bruce, Peter Gedeck, O'Reilly, 2020
  - ❖ Introduction to Machine Learning with Python, Andreas C. Müller & Sarah Guido, O'Reilly, 2019.
  - ❖ Data Science from Scratch, 2nd Edition, Joel Grus, O'Reilly, 2019
  - ❖ Essential Math for Data Science, Thomas Nield, O'Reilly, 2022
  - ❖ Python pour le Data Scientist, Emmanuel Jakobowicz, Dunod, 2018.