Rapport 2 : 6 pages

## **Introduction**

Dans le cadre du projet d’estimation des prix d’immobiliers, ce rapport représente une description détaillée de nos données ainsi que l’états d’avancement de notre étude. Des études ont été réalisées afin de trouver une méthode alternative de prédiction de prix d’immobilier, autre que les méthodes conventionnelles. Pour ce faire, plusieurs variables doivent être prises en compte d’où la pertinence de l’analyse.

Notre projet est dans le but d’améliorer l’évaluation des valeurs d’immobilier dans la ville d’Ames aux États-Unis. Notre objectif principal est la prédiction des prix de vente des immobiliers selon leur caractéristiques et leurs historiques. Pour ce faire, l’analyse et la compréhension de données est une étape primordiale. Cette analyse comprend une description détaillée des données et de leurs difficultés, une étude statistique, la correction des anomalies, le traitement des données manquantes ainsi qu’une revue de littérature.

## **Description des données :**

### Discutez des données

Dans l’optique de prédire le prix d’une maison, nous avons besoins d’avoir le maximum d’information sur la propriété afin de minimiser l’erreur de prédiction. Dans le cas présent, les données pour cette étude contiennent 79 variables pour 2919 échantillons dont 1460 immobiliers sont avec le prix d’achat connu.

Bien entendu, les variables n’ont pas tous la même importance et leur influence sur le prix de vente diffère en fonction de la corrélation avec la valeur de vente. Dans cette phase d’exploration de données, nous avons procédé avec une simple étude de corrélation de plus à une lecture logique de l’utilité des autres variables.

À cet effet, les dix variables les plus importantes avec une première analyse sont les suivantes :

|  |  |
| --- | --- |
| Variable | Corrélation |
| SalePrice | 1.000000 |
| OverallQual | 0.790982 |
| GrLivArea | 0.708624 |
| GarageCars | 0.640409 |
| GarageArea | 0.623431 |
| TotalBsmtSF | 0.613581 |
| 1stFlrSF | 0.605852 |
| FullBath | 0.560664 |
| TotRmsAbvGrd | 0.533723 |
| YearBuilt | 0.522897 |
| YearRemodAdd | 0.507101 |
| GarageYrBlt | 0.486362 |
| MasVnrArea | 0.477493 |

### Les types d’attributs

Au total, nous avons 46 variables de type entier qui forment le groupe de données quantitative, en revanche, pour les données qualitatives nous avons 35 variables de type objet. Vu du nombre élevé des variables, un résumé détaillé concernant les variables se trouve dans l’annexe A du document ci-présent.

Cependant, une étude préliminaire était faite pour avoir une meilleure compréhension sur la pertinence de ces derniers. Ainsi, les attributs se divise sous X catégories :

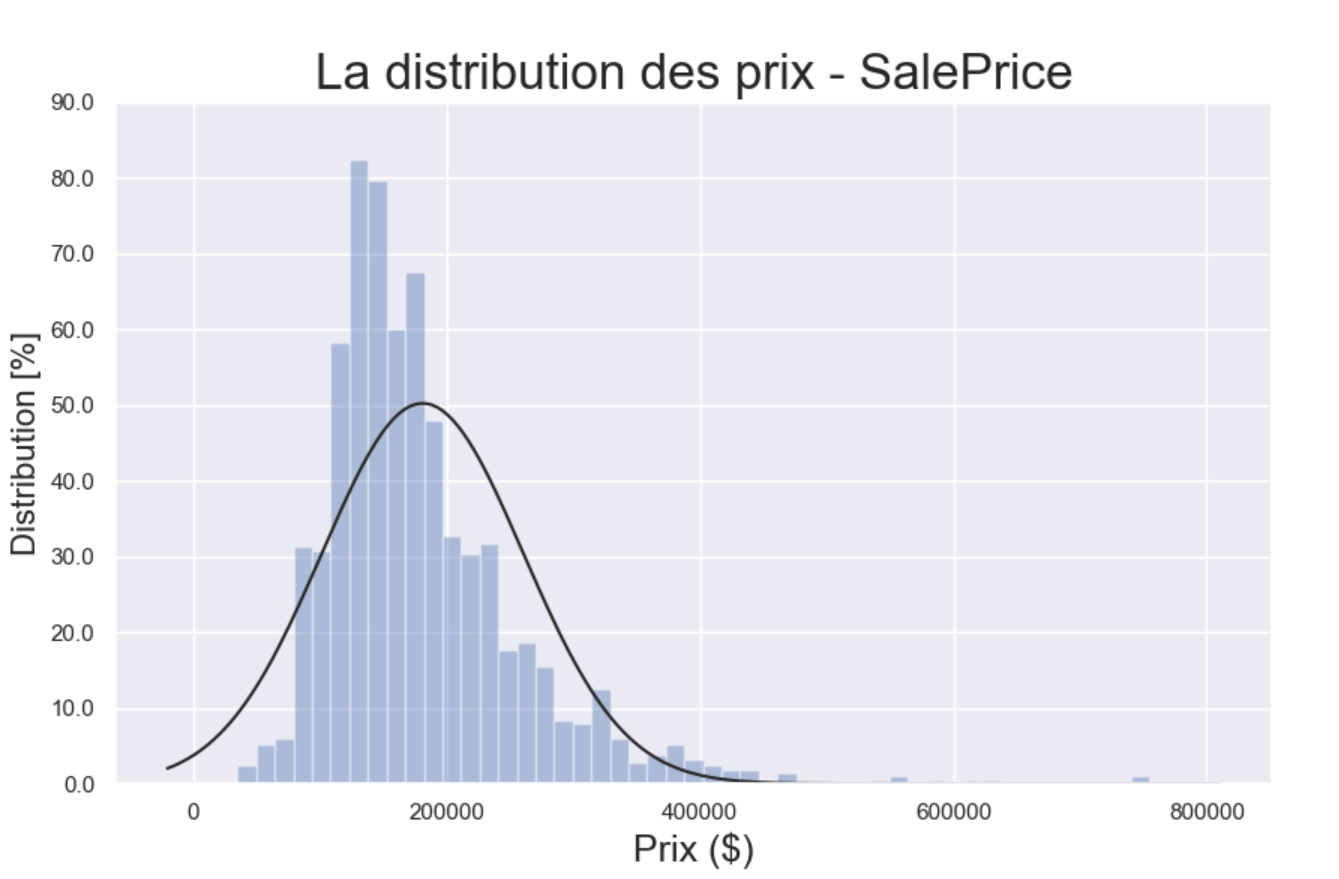
Variables temporaires : ces variables sont de type de date qui peuvent être en lien avec l’année de vente de construction ou de ventre. Exemple : YrSold, YearRemodAdd, YearBuilt, GarageYrBlt.

Variable prix : soit le prix de la maison qu’on doit prédire ou le prix de certaines caractéristiques précise. SalePrice et MiscVal.

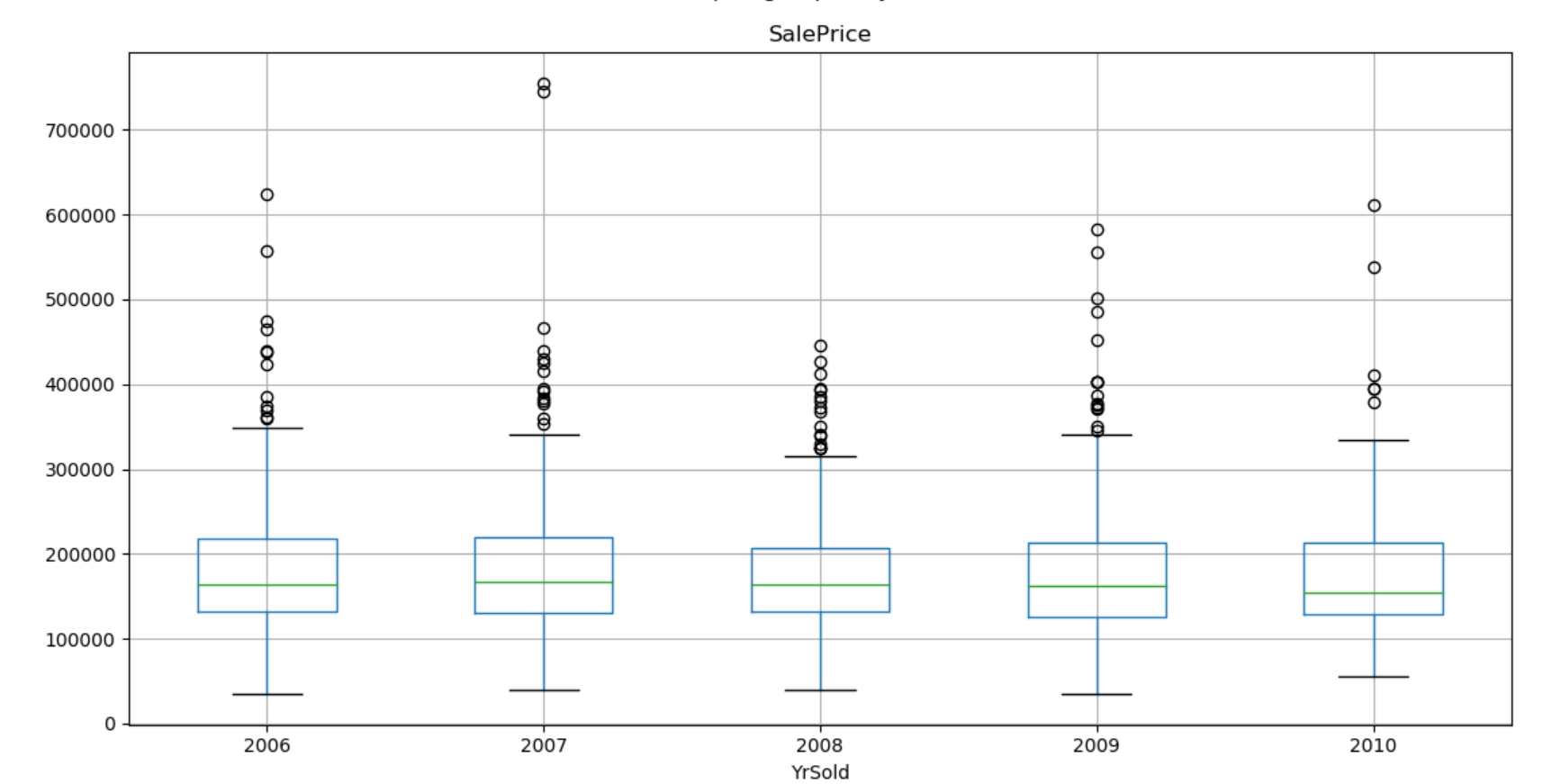
……

### Les propriétés statistiques des données.

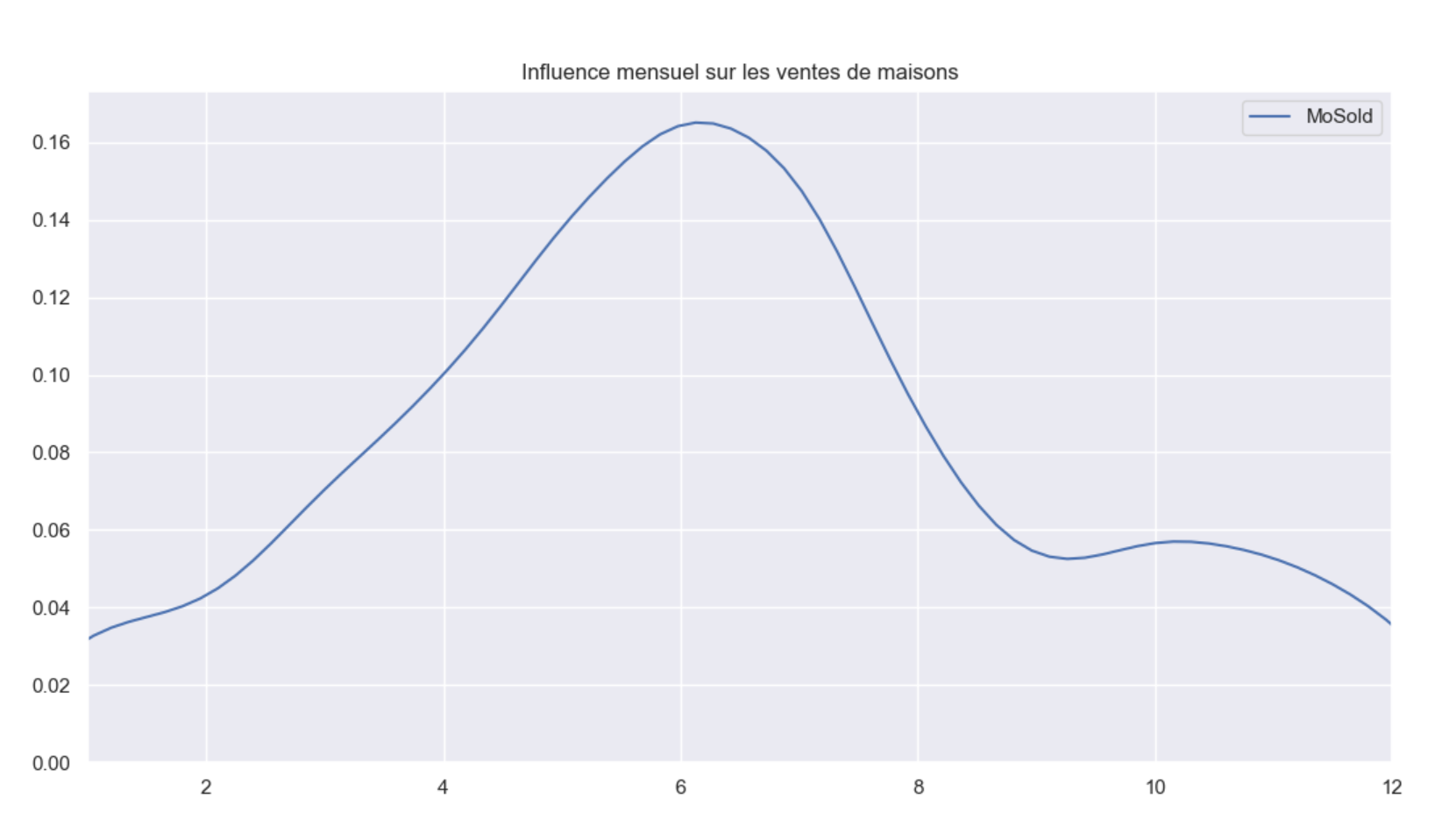
Puisque nous voulons prédire le prix des maisons dans la ville d’Ames pour les années 2006 à 2010, il est important de savoir la distribution des prix sur les 1440 maisons connues.



Nous sommes en mesure de remarquer que la majorité des prix sur l’ensemble des données (80% des maisons) ont un prix entre 120 000$ à 200 000$. De même, certaines valeurs sont aberrantes par rapport à la moyenne observée.



Si nous examinons la même variable « *SalePrice* » plus en détail, on peut déduire que les valeurs exagérées des prix de maisons loin de la médiane sont vendues plus en 2006 et 2009, aussi, deux propriétés en particulier ont étaient vendus en 2007 pour une valeur supérieure à 700 000$.

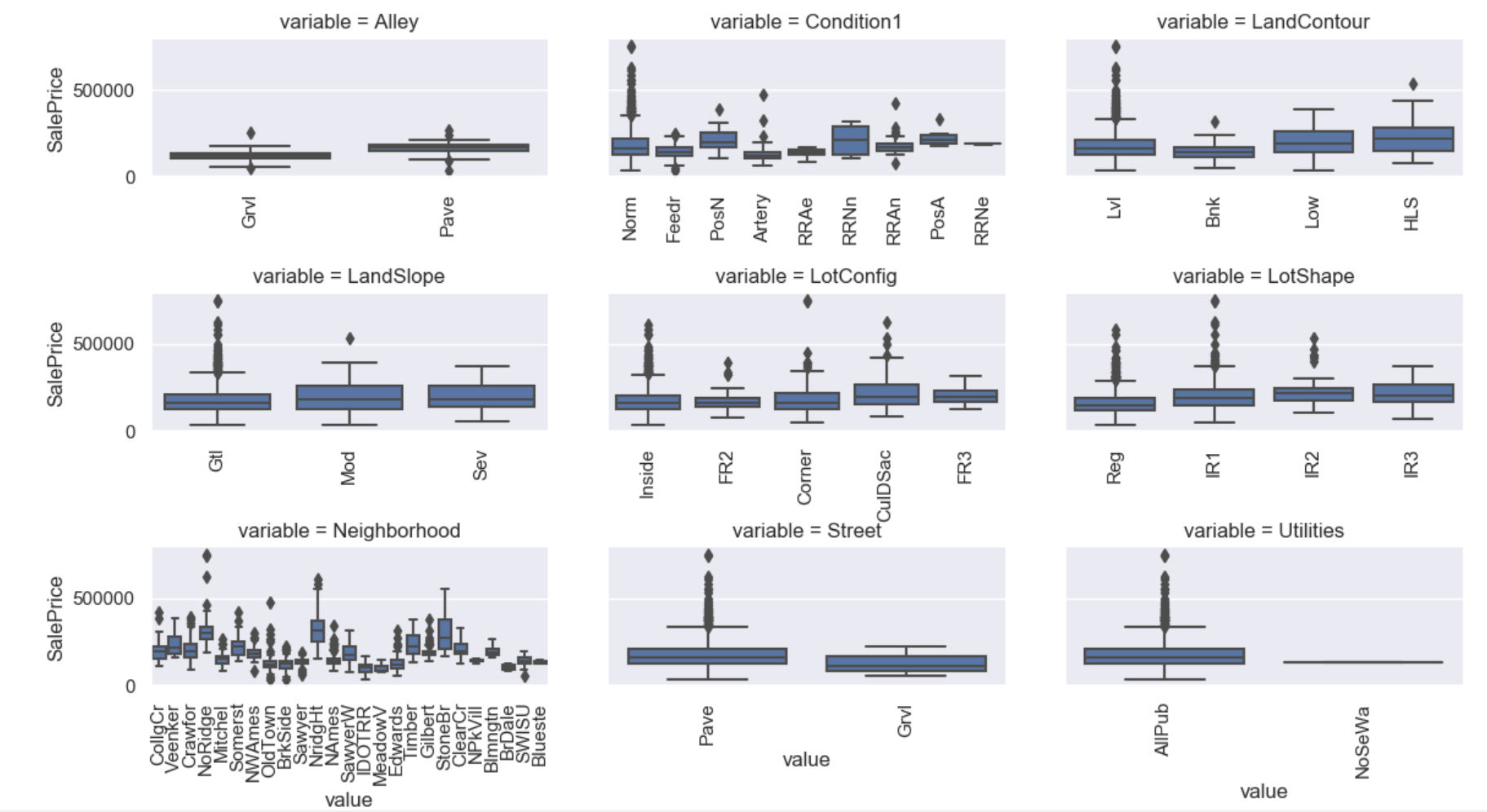


À l’intérieure d’une même année, il y a présence d’une variété sur le taux de vente. L’été étant la meilleure saison de vente et influence ainsi le prix des maisons. La variable *MoSold* nous permet de détecter ce changement qui peut aider la prédiction des ventes.

## **Difficultés et algorithmes de prétraitement (2 points)**

### Quelles difficultés se présentent?

La première difficulté pour les algorithmes de prétraitement est le type de données qu’on doit avoir en entré, ce type doit être numérique. Toutefois, nous avons plusieurs variables de type « objet », autrement dit, ils doivent être transformées en type « *categroy* ».



Évidemment, la distribution d’une variable qualitative par groupe n’est pas homogène, ceci amène du bruit sur l’ensemble des données. En se fiant sur la figure ci-haut, ce cas est plus flagrant pour les variables : *Utilities* et *Neighborhood*. La proportion d’information est très limitée dans une seule catégorie. Cette remarque amène la question suivante : Est-il important de garder cette variable ?

(ex. : bruit, fléau de dimensionnalité, informations manquantes, déséquilibre des classes, etc.).

### Décrivez ces algorithmes

Implantez des algorithmes de prétraitement des données afin de corriger ces difficultés.

### Discutions de leurs résultats

**L’objectif ici n’est pas de faire une grande liste de statistiques sur les données, mais d’en tirer des leçons pour guider la réalisation du reste du projet.**

## **Données manquantes (1 point**)

Discutez également de ce qui manque dans vos données.

Quelles informations additionnelles voyez-vous comme nécessaire ou utile à avoir pour réaliser votre projet mais manquent au jeu de données de Kaggle?

Où et comment pourriez-vous obtenir ces informations supplémentaires afin d’enrichir votre jeu de données?

Comment est-ce que ces différents jeux de données vont être combinés dans votre système?

## **Procédure de tests (1 point)**

Discutez également de la procédure de tests que vous envisagez.

La majorité des projets Kaggle ne viennent pas avec des données tests, et l’option de soumettre un fichier de résultats pour évaluation ne retourne qu’un score numérique sans indications de ce qui a ou n’a pas fonctionné. On ne peut pas faire un projet en tâtant dans le noir!

Donc, comment prévoyez-vous tester vos solutions afin d’obtenir une rétroaction qui pourra guider votre développement?

## **Revu de littérature (2 points)**

Allez chercher des idées dans les travaux antérieurs. Votre première destination devrait être le forum de discussion associé à votre compétition Kaggle, où d’autres équipes peuvent avoir échangé des idées. Vous pouvez également contacter certaines des équipes plus hautes dans le classement pour leur poser des questions.

Finalement, vous pouvez vous tourner vers la littérature scientifique, où des projets similaires ont déjà été réalisés.

Ceci vous donnera des idées sur comment traiter vos données Kaggle, et comment planifier pour les prochaines étapes. N’oubliez pas de bien inclure vos références!

Dans le but d’améliorer l’évaluation des valeurs d’immobilier, des études ont été réalisées afin de trouver une méthode alternative de prédiction de prix d’immobilier autre que les méthodes conventionnelles telle que la méthode Hedonic.

Dans l’étude comparative des modèles de prédiction des prix d’immobilier rural et urbain en Turquie, Hasan S. (2008) a démontré que le modelé basé sur les réseaux de neurones artificiels est significativement plus performante la régression de Hedonic avec une différence d’erreur quadratique (MSE) estimée à 2,03.

Selon Bourassa S. C. et al. (2007) la méthode de régression Hedonic ne prend pas en considération l’effet de la localisation sue les prix. Xiaalong L. (2012) a démontré l’importance de prise en considération des effets temporaire et spatiale sur l’estimation des prix. Ce dernier a mentionné aussi que le méthode Hedonic ne prend pas en considération les effets spatio-temporels.

## **Comparaison avec les intuitions rapport 1 (1 point)**

Finalement, faites un lien avec votre premier rapport. À la lumière de ce que vous avez appris maintenant, les idées et intuitions que vous aviez précédemment sont-elles encore valides? Si oui,

Laquelle prévoyez-vous implémenter en premier, et pourquoi semble-t-elle la plus prometteuse? Si non, comment allez-vous les mettre à jour?

# **Bibliographie**

Bourassa, S. C., Cantoni, E., & Hoesli, M. (2007). Spatial dependence, housing submarkets, and house price prediction. *The Journal of Real Estate Finance and Economics*, *35*(2), 143-160.

Bourassa, S., Cantoni, E., & Hoesli, M. (2010). Predicting house prices with spatial dependence: a comparison of alternative methods. *Journal of Real Estate Research*, *32*(2), 139-159.

Limsombunchai, V. (2004, June). House price prediction: hedonic price model vs. artificial neural network. In *New Zealand Agricultural and Resource Economics Society Conference* (pp. 25-26).

Liu, X. (2013). Spatial and temporal dependence in house price prediction. *The Journal of Real Estate Finance and Economics*, *47*(2), 341-369.

Selim, H. (2009). Determinants of house prices in Turkey: Hedonic regression versus artificial neural network. *Expert Systems with Applications*, *36*(2), 2843-2852.

Annexe A

Liste des variables utilisé dans la phase de préparation des données qui servent à la prédiction du prix d’une maison.

