Hiver 2019



Rapport 2 : Données du projet d'estimation des prix d'immobilier

GLO-7027 Analyse et traitement de données massives

Ali ASSAFIRI 111 054 128

Rhita OULIZ 111 082 917

14 février 2019

Table des matières

[**Introduction** 3](#_Toc1504246)

[**Description des données** 3](#_Toc1504247)

[**Les données** 3](#_Toc1504248)

[**Les types d’attributs** 3](#_Toc1504249)

[**Les propriétés statistiques des données** 5](#_Toc1504250)

[**Difficultés et algorithmes de prétraitement** 6](#_Toc1504251)

[**Les difficultés au niveau des données** 6](#_Toc1504252)

[**Données manquantes** 7](#_Toc1504253)

[**Procédure de tests** 8](#_Toc1504254)

[**Revu de littérature** 8](#_Toc1504255)

[**Comparaison avec les intuitions rapport 1 (1 point)** 9](#_Toc1504256)

[**Bibliographie** 10](#_Toc1504257)

[Figure 1 - Illustration de la matrice de distance entre les variables numériques du jeu de données 4](#_Toc1504120)

[Figure 2 - la distribution des prix de vente des immobiliers. 5](file:///C:\Users\ali_a\Desktop\HIV-2019\GLO-7029_AnalyseBigData\GLO7029_01_H19-HousePricing\Rappor\Rapport%202%20-%20Données%20V5.docx#_Toc1504121)

[Figure 3 - la distribution des prix de vente des immobiliers selon les années. 6](#_Toc1504122)

[Figure 4 - Influence saisonnière sur le taux de vente. 14](#_Toc1504123)

[Tableau 1 - Nombre d’attributs par type. 4](#_Toc1504282)

[Tableau 2 - les variables les plus corrélées avec le prix de vente d’immobilier. 7](#_Toc1504283)

## **Introduction**

Dans le cadre du projet d’estimation des prix d’immobiliers, ce rapport représente une description détaillée de nos données ainsi que l’état d’avancement de notre étude.

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 caractéristique et leurs historiques. Afin de comprendre les données de ce projet, nous allons élaborer 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**

### **Les données**

Dans l’optique de prédire le prix d’une maison, nous avons besoin 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.

### **Les types d’attributs**

Au total, nous avons 46 variables de type entier qui forment le groupe de données quantitatives. 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 divisent sous 3 catégories :

**Variables nominales** :Il s’agit des attributs descriptifs tel que le nom des rues, le type de la vente ou autres. Ce type est le plus omniprésent dans notre base de données avec un taux de 42 sur 79 attributs.

**Variable ordinale**: Ce sont les attributs d’évaluation de l’état général des éléments de l’immeuble comme la qualité générale du matériau et de la finition : OverallQual.

**Variables numériques :**

En se basent sur une étude de distance entre les variables numériques, nous sommes en mesure de les regrouper sous 4 groupes, comme illustrés dans la figure 1.

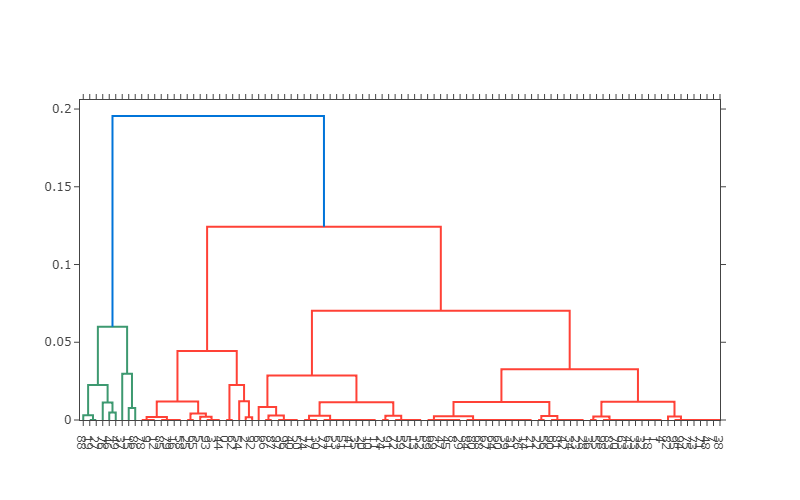


Figure 1 - Illustration de la matrice de distance entre les variables numériques du jeu de données

Variables numériques à l’échelle d’intervalle: 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.

Variables numériques à l’échelle de ratio: ces variables représentent l’aire en pieds carrés du terrain de l’immeuble, du garage et du sous-sol et autre espace (GrLivArea, GarageArea, MasVnrArea).

Variable continue (le prix) : soit le prix de la maison qu’on doit prédire ou le prix de certaines caractéristiques précises telles que SalePrice et MiscVal.

Variables discrètes : précise le nombre des pièces à chaque que soit le nombre de chambres, le nombre des salles de bain ou le nombre des cuisines.

|  |  |
| --- | --- |
| Type d'attribut | Nombre d'attributs |
| Nominal | 42 |
| Numérique à l'échelle de ratio | 18 |
| Numérique à l'échelle d'intervalle | 5 |
| Numérique continu | 1 |
| Numérique discret | 10 |
| Ordinal | 4 |

Tableau 1 - Nombre d’attributs par type.

### **Les propriétés statistiques des données**

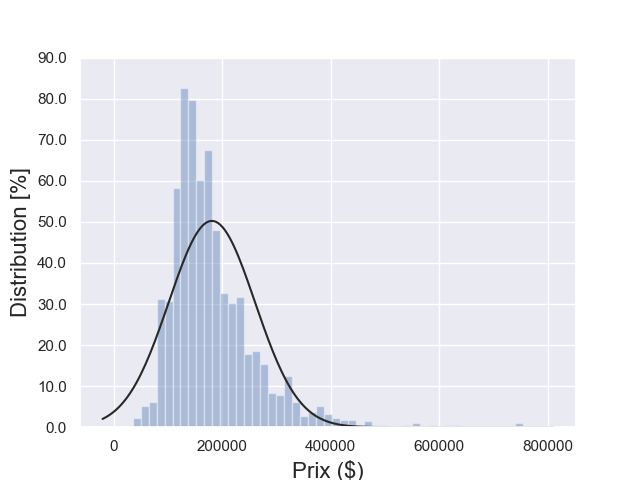
Puisque nous voulons prédire le prix des maisons dans la ville d’Ames pour les années 2006 à 2010, il est important de savoir la distribution des prix sur les 1440 maisons connues.

Figure - la distribution des prix de vente des immobiliers.

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.

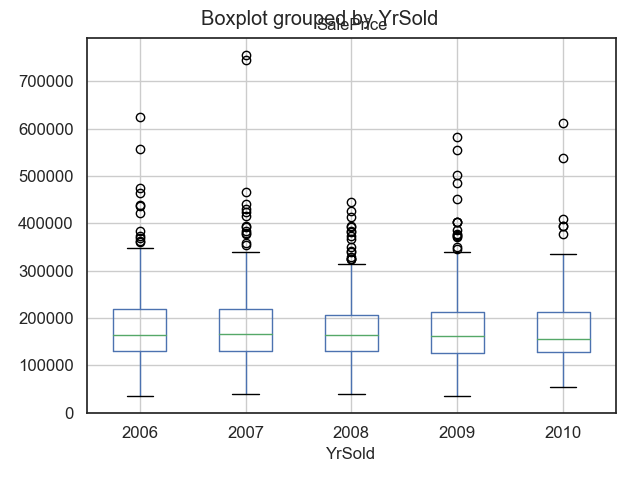


Figure 3 - la distribution des prix de vente des immobiliers selon les années.

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 était vendu en 2007 pour une valeur supérieure à 700 000$. Il est de même important de noter que les ventes sont plus nombreuses durant l’été par rapport aux autres saisons comme l’illustre la figure 5 dans l’annexe B.

## **Difficultés et algorithmes de prétraitement**

### **Les difficultés au niveau des données**

La première difficulté pour les algorithmes de prétraitement est le type de données qu’on doit avoir en entrée, ce type doit être numérique. Toutefois, nous avons plusieurs variables de type « objet », autrement dit, ils doivent être transformés 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. Avec 79 variables, la dimensionnalité présente ainsi un défi.

De plus, comme mentionnées dans la section précédente, certaines variables contiennent des valeurs aberrantes et très loin de la moyenne. Ces valeurs en général influencent négativement les statistiques qui servent à remplir les valeurs manquantes de la même variable.

Une autre problématique remarquée est au niveau de la pertinence des catégories présentes pour une même variable. À titre d’exemple, pour la variable 1stFlw

**Les algorithmes de traitement de données**

Afin de corriger ces difficultés, nous avons implanté des algorithmes de prétraitement des données :

* Le premier algorithme est un formateur de données qui trie les données en fonction de leurs types de valeurs. S’il détecte un type « objet », il transforme ce dernier en type « category », il rend les valeurs de chaque catégorie en entier qui lui correspond tout en gardant un index pour sauvegarder les assignations.
* Le deuxième algorithme est de rendre les catégories d’une seule variable en colonnes, ces colonnes s’ajout sur la base de données initiale sous la forme de valeur booléenne. À titre d’exemple, la variable *HouseType* à 5 catégories : *Abnormal, Normal, AdjLand, Family* et *Partial*. À l’aide de l’algorithme en question cette colonne : *HouseType* devient 5 colonnes où chaque colonne porte le nom de sa catégorie.

* Le troisième algorithme sert à remplacer des valeurs inconnues « NaN » ou vides avec les statistiques de la même variable. En général, nous avons le choix de remplacer les valeurs vides par la moyenne ou la médiane.

**Les résultats du traitement de données**

Après le traitement des données des variables nominales et le filtrage des variables qui bruitent les données, une étude de corrélation a été réalisée.

Bien entendu, les variables n’ont pas toutes 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 | Description | Corrélation |
| OverallQual | Qualité générale du matériau et de la finition | 0,79 |
| GrLivArea | Surface habitable au-dessus du niveau du sol | 0,71 |
| GarageCars | Taille du garage en capacité de la voiture | 0,64 |
| GarageArea | Taille du garage en pieds carrés | 0,62 |
| TotalBsmtSF | Nombre total de pieds carrés de sous-sol | 0,61 |
| 1stFlrSF | Premier étage pieds carrés | 0,61 |
| FullBath | Salles de bain complètes au-dessus du niveau | 0,56 |
| TotRmsAbvGrd | Nombre total de chambres au-dessus du sol | 0,53 |
| YearBuilt | Date de construction originale | 0,52 |
| YearRemodAdd | Date de remodelage | 0,51 |
| GarageYrBlt | Année de construction du garage | 0,49 |
| MasVnrArea | Surface de placage de maçonnerie en pieds carrés | 0,48 |

Tableau 2 - les variables les plus corrélées avec le prix de vente d’immobilier.

## **Données manquantes**

Dans le cas de notre base de données, les variables avec des valeurs nulles ne sont pas nombreuses. L’annexe B représente le nombre de valeurs non nulles pour chaque attribut.

Toutefois, la crise économique qui a eu lieu en 2009 nous mené a évalué son influence sur les prix d’immobilier. Selon la figure 2 ci-dessus, la distribution des prix de vente en 2009 n’est pas très différente de celles des autres années. Ce qui nous mène à chercher de l’information supplémentaire pour enrichir notre étude.

De plus, les prix de vente d’immobilier sont influencés par les effets spatiotemporels, selon la littérature. Alors que dans notre étude de corrélation les attributs de localisation ne sont pas significativement corrélés avec le prix de vente.

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

Pour remédier à ces deux problématiques, nous sommes allés chercher de l’information complémentaire qui peut servir à compléter la base de données déjà en possession. Ces deux nouvelles bases de données représentent la densité de la population par cartier pour la Ville d’Ames en Iowa, de plus, à l’évaluation foncière de l’état d’Iowa pour les maisons da la même ville pour la période de 2006 à 2010. La source de ces données est bien le portail gouvernemental américain et les données sont de type open source.

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

Ces deux nouvelles informations peuvent être utiles en tant que nouvelles variables indépendantes. Par contre la valeur de vente estimée par l’état peut servir comme une valeur de correction pour ajuster la différence de prix que nous remarquons en 2009.

Le choix sera fait suite à une étude plus détaillée qui quantifierait l’apport de ces deux nouvelles données sur la base de données initiale.

## **Procédure de tests**

L’évaluation de la performance de notre estimateur sera effectuée via la validation croisée.

Il s’agit d’un algorithme d’estimation des erreurs très utiliser surtout quand il s’agit des études avec un nombre d’exemples limité.

À cette étude, nous disposons de 1460 échantillons étiquetés c’est-à-dire avec le prix de vente connu. Ce qui élimine la méthode de division des données en sous-ensembles de traitement, de validation et de test. La validation croisée est alors une procédure de tests de la performance de notre estimateur adéquate pour notre cas d’étude. Cette méthode consiste à générer des jeux de données aléatoires de notre base de données une multitude de fois afin de tester leurs résultats.

## **Revue de littérature**

Les idées dans les travaux antérieurs.

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 performant 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 spatial sur l’estimation des prix. Ce dernier a mentionné aussi que le méthode Hedonic ne prend pas en considération les effets spatiotemporels.

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

Dans le premier rapport, nous avons prévu d’utiliser les réseaux de neurones comme algorithme de prédiction. Selon cette étude préliminaire des données, les nombres d’échantillons que nous avant ne permet pas de tirer profil des performances de cette méthode. Les réseaux de neurones sont plus efficaces quand il s’agit des bases de données avec un nombre d’exemples de l’ordre de 10 000.

Ce qui nous mène à éliminer l’utilisation des réseaux de neurones de notre plan d’étude (précisément de la partie d’étude comparative des résultats des deux algorithmes) et comparer plutôt les méthodes *Random Forest* et *Xgboost*.

# **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 utilisées dans la phase de préparation des données qui servent à la prédiction du prix d’une maison.

|  |  |  |
| --- | --- | --- |
| Attribut | Description | Types d'attribut |
| MSSubClass | La classe de construction | Ordinal |
| MSZoning | La classification générale de zonage | Nominal |
| LotFrontage | Pieds linéaires de la rue reliés à la propriété | Numérique à l'échelle de ratio |
| LotArea | Taille du terrain en pieds carrés | Numérique à l'échelle de ratio |
| Street | Type d'accès routier | Nominal |
| Alley | Type d'accès aux allées | Nominal |
| LotShape | Forme générale de la propriété | Nominal |
| LandContour | Planéité de la propriété | Nominal |
| Utilities | Type d'utilitaires disponibles | Nominal |
| LotConfig | Configuration du lot | Nominal |
| LandSlope | Pente de la propriété | Nominal |
| Neighborhood | Emplacements physiques dans les limites de la ville d'Ames | nominales |
| Condition1 | Proximité de la route principale ou du chemin de fer | Nominal |
| Condition2 | Proximité de la route principale ou du chemin de fer | Nominal |
| BldgType | Type de logement | Nominal |
| HouseStyle | Style d'habitation | Nominal |
| OverallQual | Qualité générale du matériau et de la finition | Ordinal |
| OverallCond | Note globale de l'état | Ordinal |
| YearBuilt | Date de construction originale | Numérique à l'échelle d'intervalle |
| YearRemodAdd | Date de remodelage | Numérique à l'échelle d'intervalle |
| RoofStyle | Type de toit | Nominal |
| RoofMatl | Matériau de toiture | Nominal |
| Exterior1st | Revêtement extérieur sur la maison | Nominal |
| Exterior2nd | Revêtement extérieur de la maison (si plus d'un matériau) | Nominal |
| MasVnrType | Type de placage de maçonnerie | Nominal |
| MasVnrArea | Surface de placage de maçonnerie en pieds carrés | Numérique à l'échelle de ratio |
| ExterQual | Qualité des matériaux extérieurs | Nominal |
| ExterCond | État actuel du matériau à l'extérieur | Nominal |
| Foundation | Type de fondation | Nominal |
| BsmtQual | Hauteur du sous-sol | Nominal |
| BsmtCond | État général du sous-sol | Nominal |
| BsmtExposure | Murs de sous-sol de plain-pied ou de jardin | Nominal |
| BsmtFinType1 | Qualité de la zone finie du sous-sol | Nominal |
| BsmtFinSF1 | Type 1 fini pieds carrés | Numérique à l'échelle de ratio |
| BsmtFinType2 | Qualité de la deuxième zone finie (si présente) | Nominal |
| BsmtFinSF2 | Type 2 fini pieds carrés | Numérique à l'échelle de ratio |
| BsmtUnfSF | Pieds carrés inachevés du sous-sol | Numérique à l'échelle de ratio |
| TotalBsmtSF | Nombre total de pieds carrés de sous-sol | Numérique à l'échelle de ratio |
| Heating | Type de chauffage | Nominal |
| HeatingQC | Qualité et état de chauffage | Nominal |
| CentralAir | Climatisation centrale | Nominal |
| Electrical | Système électrique | Nominal |
| 1stFlrSF | Premier étage pieds carrés | Numérique à l'échelle de ratio |
| 2ndFlrSF | Pieds carrés au deuxième étage | Numérique à l'échelle de ratio |
| LowQualFinSF | Pieds carrés finis de qualité médiocre (tous les étages) | Numérique à l'échelle de ratio |
| GrLivArea | Surface habitable au-dessus du niveau du sol | Numérique à l'échelle de ratio |
| BsmtFullBath | Salle de bain complète au sous-sol | Numérique discret |
| BsmtHalfBath | Demi-salles de bain | Numérique discret |
| FullBath | Salles de bain complètes au-dessus du niveau | Numérique discret |
| HalfBath | Demi-bains au-dessus du niveau | Numérique discret |
| Bedroom | Nombre de chambres au-dessus du sous-sol | Numérique discret |
| Kitchen | Nombre de cuisines | Numérique discret |
| KitchenQual | Qualité de la cuisine | Ordinal |
| TotRmsAbvGrd | Nombre total de chambres au-dessus du sol | Numérique discret |
| Functional | Évaluation de la fonctionnalité d'accueil | Nominal |
| Fireplaces | Nombre de cheminées | Numérique discret |
| FireplaceQu | Qualité de la cheminée | Nominal |
| GarageType | Emplacement du garage | Nominal |
| GarageYrBlt | Année de construction du garage | Numérique à l'échelle d'intervalle |
| GarageFinish | Finition intérieure du garage | Nominal |
| GarageCars | Taille du garage en capacité de la voiture | Numérique discret |
| GarageArea | Taille du garage en pieds carrés | Numérique à l'échelle de ratio |
| GarageQual | Qualité de garage | Nominal |
| GarageCond | Etat du garage | Nominal |
| PavedDrive | Allée pavée | Nominal |
| WoodDeckSF | Surface de pont en bois en pieds carrés | Numérique à l'échelle de ratio |
| OpenPorchSF | Porche ouvert en pieds carrés | Numérique à l'échelle de ratio |
| EnclosedPorch | Porche fermé en pieds carrés | Numérique à l'échelle de ratio |
| 3SsnPorch | Porche trois saisons en pieds carrés | Numérique à l'échelle de ratio |
| ScreenPorch | Espace porche d'écran en pieds carrés | Numérique à l'échelle de ratio |
| PoolArea | Espace piscine en pieds carrés | Numérique à l'échelle de ratio |
| PoolQC | Qualité de la piscine | Nominal |
| Fence | Qualité de clôture | Nominal |
| MiscFeature | Autre caractéristique non couverte dans les autres catégories | Nominal |
| MiscVal | Valeur de la fonction diverse | Numérique discret |
| MoSold | Mois vendu | Numérique à l'échelle d'intervalle |
| YrSold | Ans | Numérique à l'échelle d'intervalle |
| SaleType | Type de vente | Nominal |
| SaleCondition | Condition de vente | Nominal |
| SalePrice | Prix de vente | Numérique continu |

Annexe B

Graphe illustrant l’effet saisonnier sur le nombre de vente.

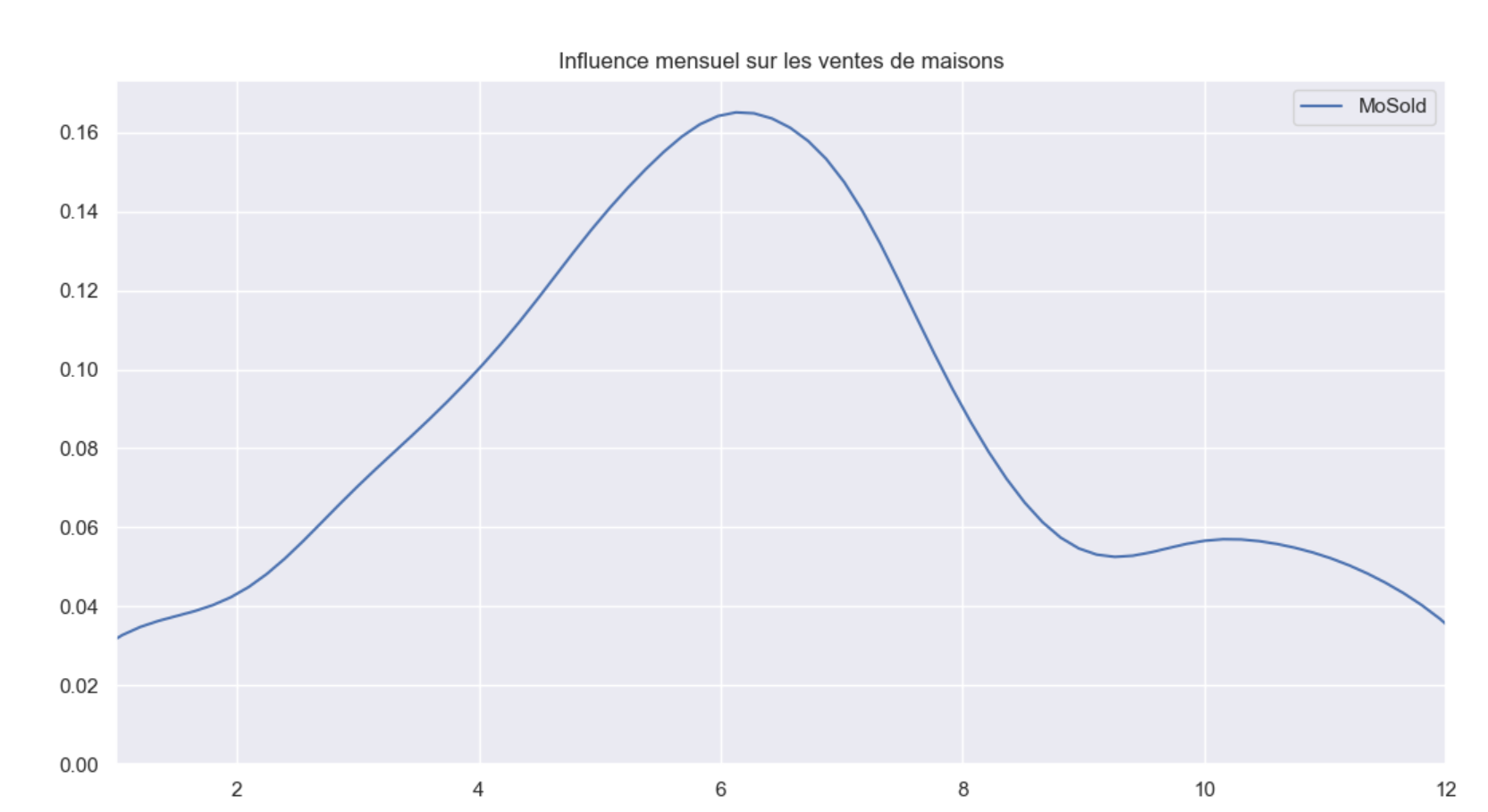


Figure 4 - Influence saisonnière sur le taux de vente.

Tableau de nombre de valeurs non nulles pour chaque variable.