|  |
| --- |
|  |

UNIVERSITÉ DE PARIS

FACULTÉ DE DROIT, ECONOMIE ET GESTION

**Facteurs expliquant le prix de vente d’une maison.**

**Prix et caractéristiques des maisons d’Ames, Iowa.**

Grégory BOURNASSENKO

Romain WU

Axel ATHOR

Groupe de TD : N° 4

Chargé de TD : M. Dondjio

LICENCE 3ÈME ANNÉE – Sciences Économiques et de Gestion

Semestre 1

**Introduction à l’économétrie**

Année universitaire 2021-2022

**1. Introduction**

Nous travaillons sur une base de données [1] qui contient le prix de vente d’un certain nombre de maisons (notre future variable expliquée) dans la ville d’Ames, dans l’Iowa, ainsi qu’un certain nombre de variables intuitivement explicatives (par exemple la surface habitable, le nombre de chambres etc.).

Nous nous intéressons donc aux facteurs qui peuvent expliquer le prix d’un bien immobilier, bien que l’analyse porte uniquement sur une ville aux USA, et donc les conclusions de cette analyse seront vraies surtout pour cette ville aux Etats-Unis en particulier.

Nous partons du principe que chaque maison a des points semblables mais aussi des différences, et se sont ces différences qui vont induire des changements sur le prix de vente. Ces différences peuvent être inhérentes aux maisons (par exemple les dimensions, le nombre de chambres, etc.) mais aussi externes (notamment l’appartenance à un quartier).

Selon les statistiques du gouvernement américain [2], le coût du logement constitue la dépense la plus importante pour la plupart des ménages. Une étude des facteurs influençant la détermination du prix des maisons pourrait aider les consommateurs à prendre des décisions importantes sur ce dont ils ont besoin ou ce qu'ils veulent dans une maison par rapport à ce qu'ils sont prêts à dépenser. Une analyse de ces facteurs pourrait être intéressante et pourrait fournir des résultats intuitifs ou parfois inattendus.

Nous nous poserons donc la question suivante :

Quels sont les facteurs qui influent le plus significativement sur la détermination du prix de vente d’une maison ?

Dans un premier temps, nous présenterons les données sur lesquelles nous basons notre analyse.

Dans un second temps, nous proposerons différents modèles en expliquant les raisons de ces choix.

En troisième partie, nous présenterons et interpréterons les résultats de ces modèles.

Pour conclure, nous résumerons notre analyse, en précisant ces avantages et inconvénients.

**2. Données**

2.1. Sources et échantillon

Notre base de données est connue et facilement accessible de plusieurs manières :

* Depuis le package “modeldata”, importer le dataset “ames”
* Depuis Github :<https://github.com/topepo/AmesHousing>)
* Depuis Kaggle: <https://www.kaggle.com/c/house-prices-advanced-regression-techniques>

Cette base de données a été compilé par [Dean De Cock](https://www.truman.edu/faculty-staff/decock/) pour être utilisée dans le cadre d’un enseignement des sciences des données. Il s'agit d'une alternative pour les spécialistes des données qui recherchent une version modernisée et étendue de l'ensemble de données sur le logement de Boston, souvent cité.

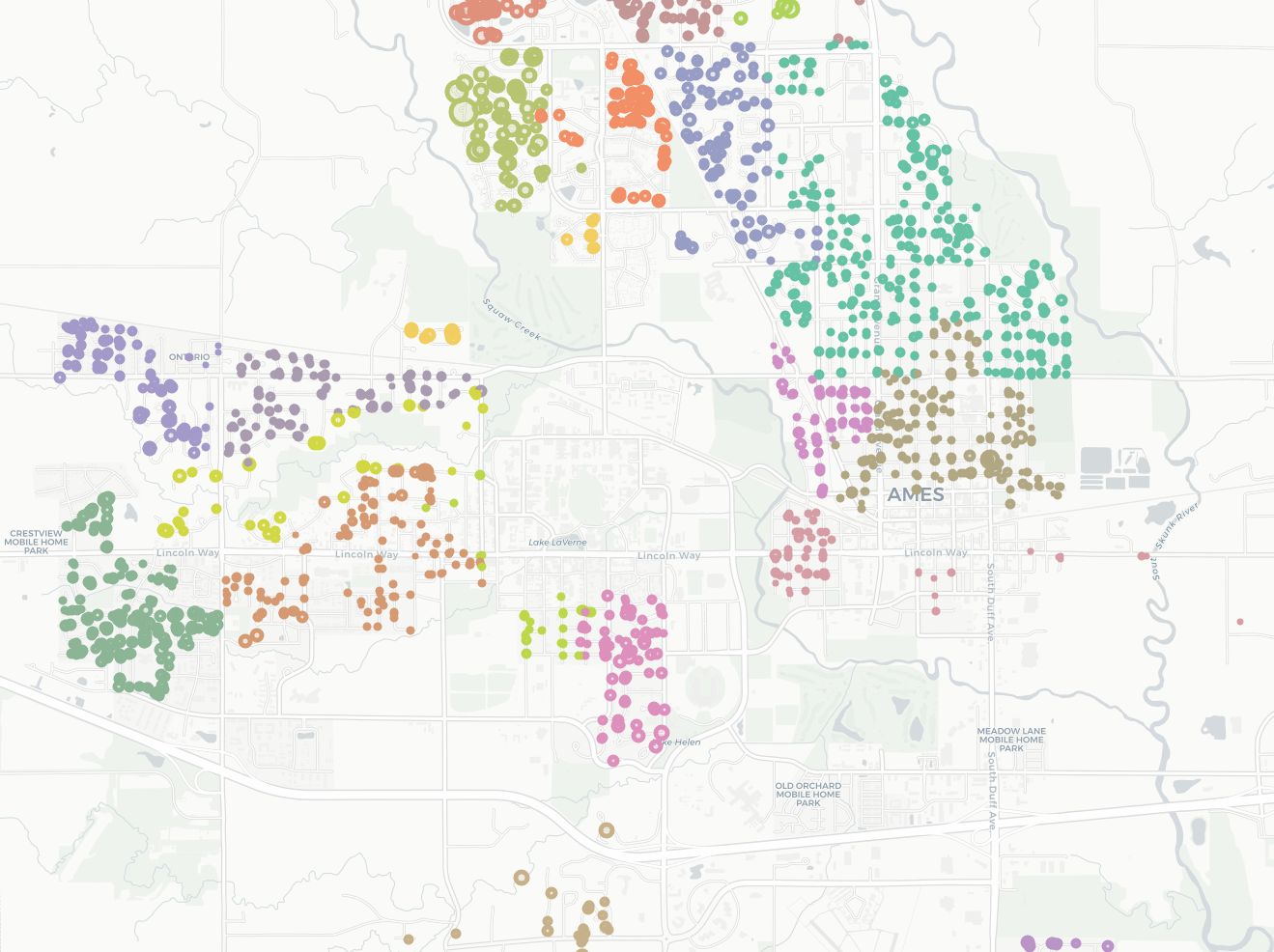
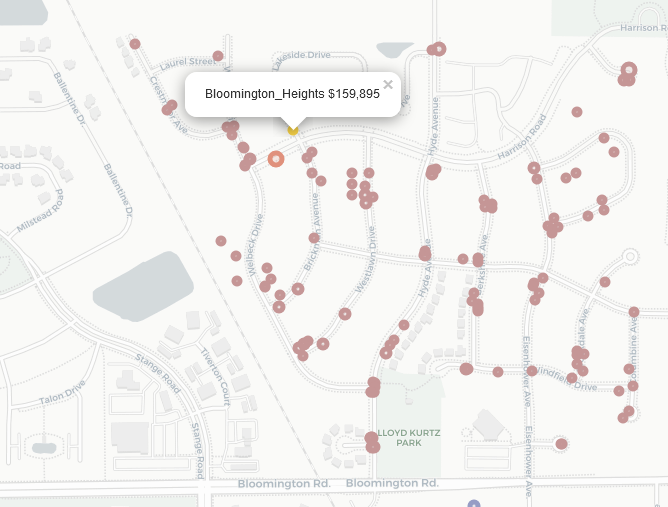
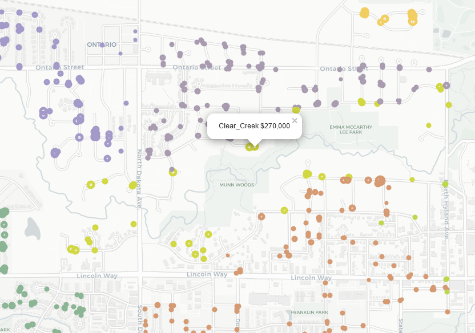
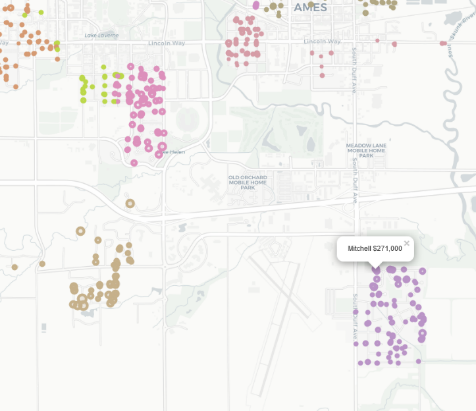
Il s’agit donc d’un grand échantillon de données relatives aux maisons, dispersées dans plusieurs quartiers de la ville d’Ames. Les dimensions (initiales) de cet échantillon sont de 2930 lignes pour 74 colonnes. Attention à ne pas confondre Ames dans l’Iowa (66 000 hab.) et Ames dans le Pas-de-Calais (652 hab.).

Avec 74 variables, nous allons devoir nous séparer de certaines, pour nous intéresser aux plus pertinentes. Cette pertinence sera définie de plusieurs sortes au fil de la recherche. La définition précise de chaque variable (en anglais) est disponible via ce [lien](http://jse.amstat.org/v19n3/decock/DataDocumentation.txt).

Notre base de données ne contient initialement aucune donnée manquante. Elle est constituée d’environ 50% de variables continues et 50% de variables discrètes.

Nous allons cependant procéder à un peu de nettoyage de données. En effet, il existe plusieurs types de ventes (Normal, Abnormal, AdjLand…) et plusieurs types d’habitation (1Fam, 2Fam, Duplx…).

Par souci de simplicité, nous travaillerons uniquement avec les ventes normales et les habitations à une (1) famille. Une fois ce nettoyage fait, nous nous retrouvons avec une base de données de dimension 2 002\*74, soit 2 002 observations et 74 variables. Voici un aperçu de la localisation des habitations par quartier (couleurs) et par prix (diamètre des cercles) :



Il y a des quartiers assez isolés des autres

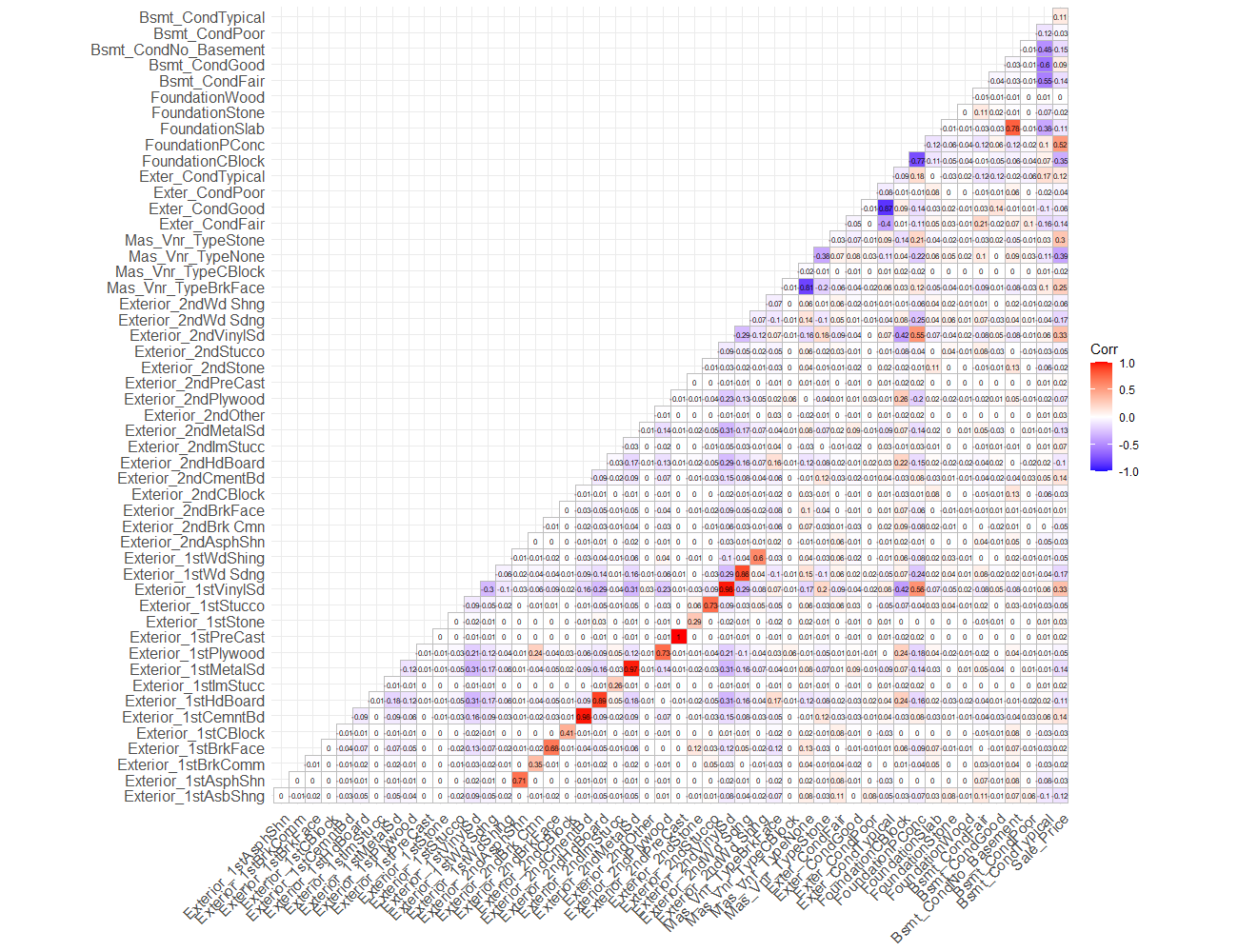
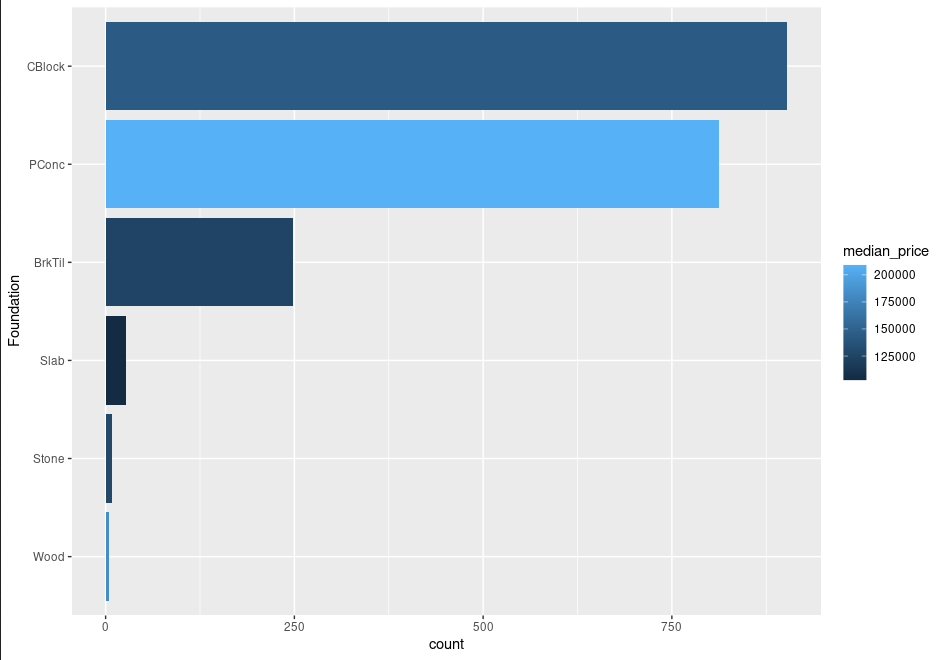
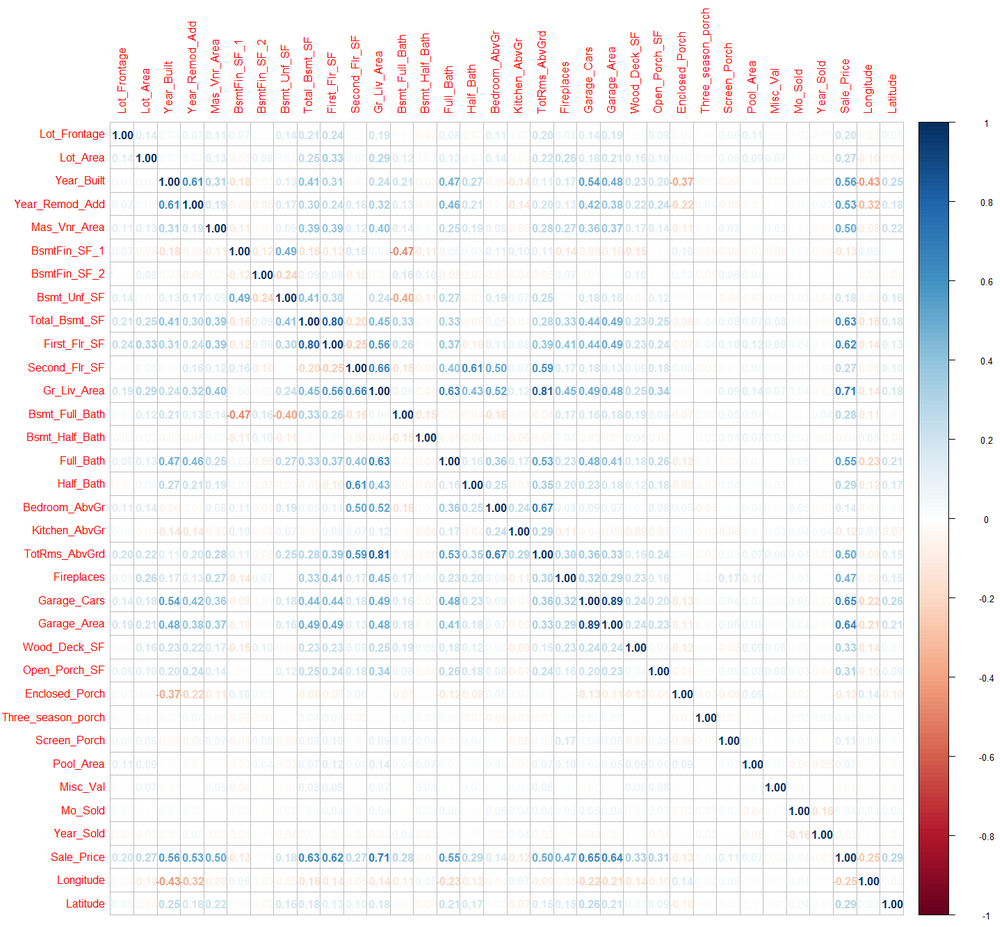
Il y a des quartiers qui sont mal délimités

Il y a une maison qui représente tout un quartier (surement à cause du nettoyage de données)

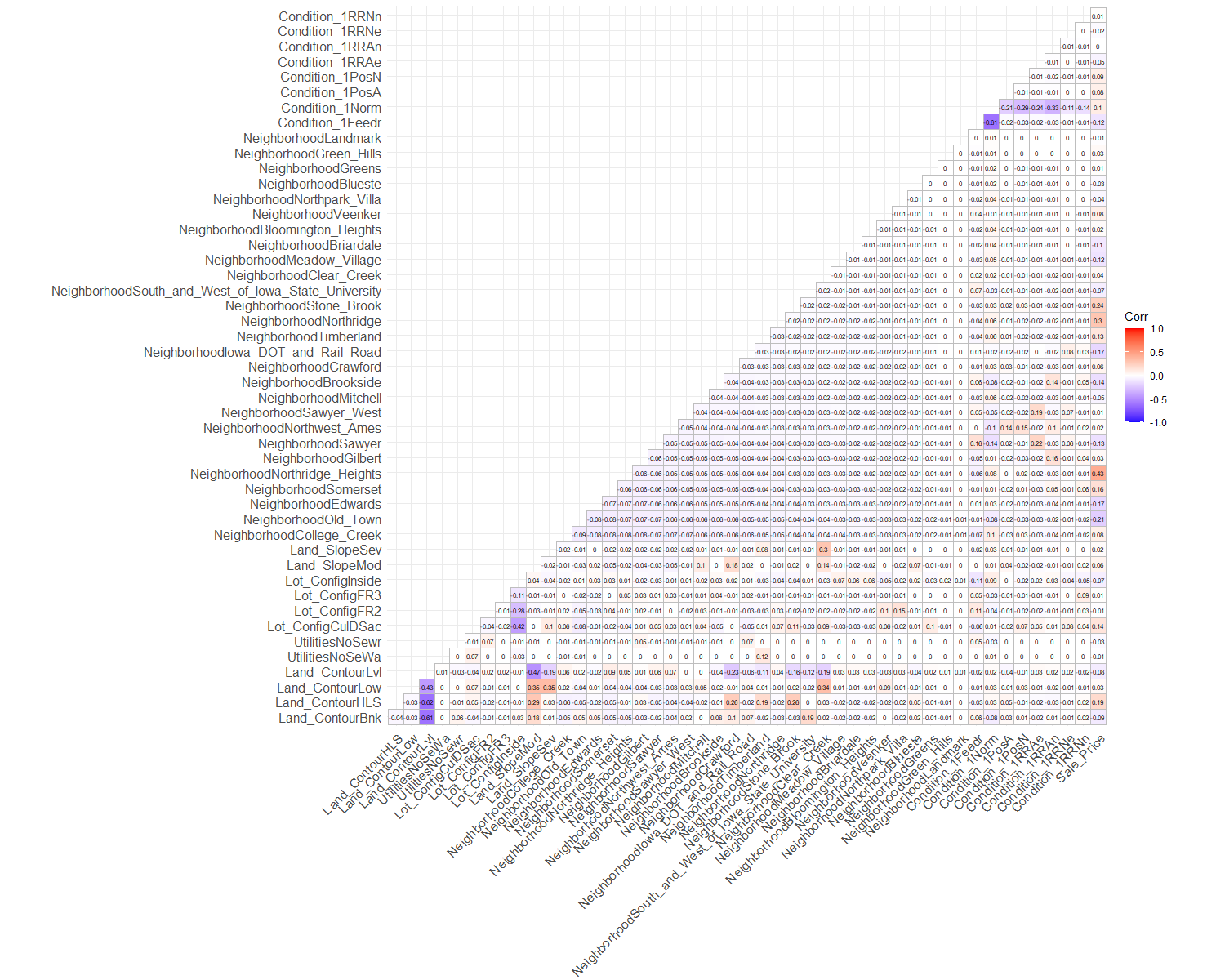
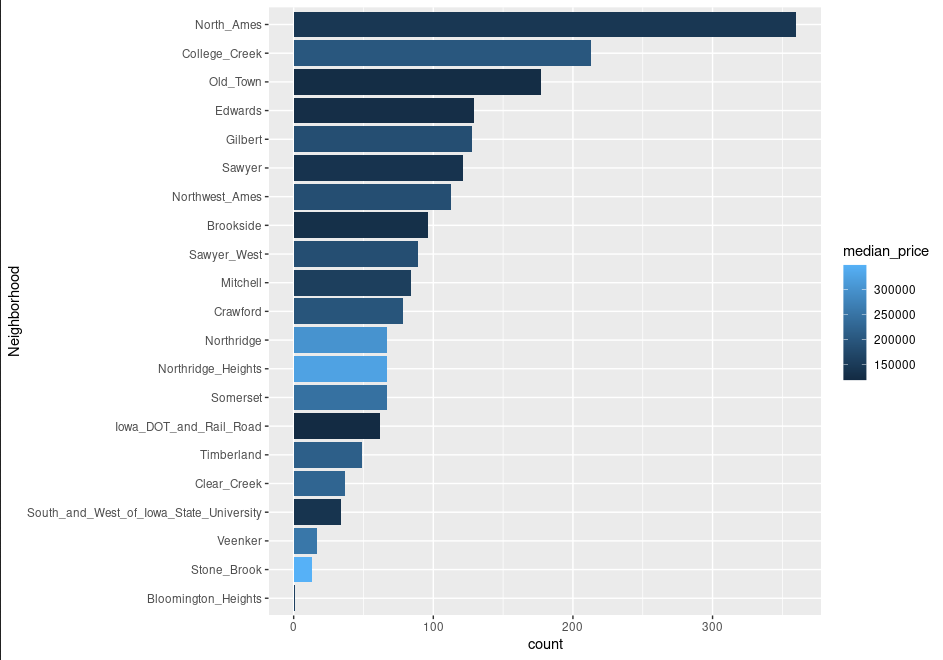
Il n’y a pas de données au centre de la ville car c’est l’emplacement d’une université

2.2. Variables

À la suite du nettoyage de données, nous avons violé l’hypothèse de non-nullité de la variance de la variable indépendante pour les variables «Building Type» ( = 1Fam) et « Sale Condition» ( = Normal), puisqu’elles sont maintenant constantes). Nous allons devoir les supprimer (et cela n’affecte en rien les futures modèles).

Choisir des variables à ce stade est assez délicat puisqu’il y a 74 variables et lorsque l’on régresse le logarithme du prix de vente sur toutes les (74-2-1)71 variables, , dans ce modèle, donc la marge de progression pour extraire du terme d’erreur des variables qui pourraient mieux expliquer log(Sale\_Price) est moindre. En raison du grand nombre de variables, nous ne présenterons que celles dont la corrélation avec le prix de vente est supérieure à 50% en valeur absolue : On remarque qu’il y a une corrélation supérieure à 50% avec le type de fondation, notamment lorsque les fondations sont en béton coulé :

La plupart des maisons vendues sont en CBlock (parpaing), mais les plus chères sont en PConc (béton coulé).

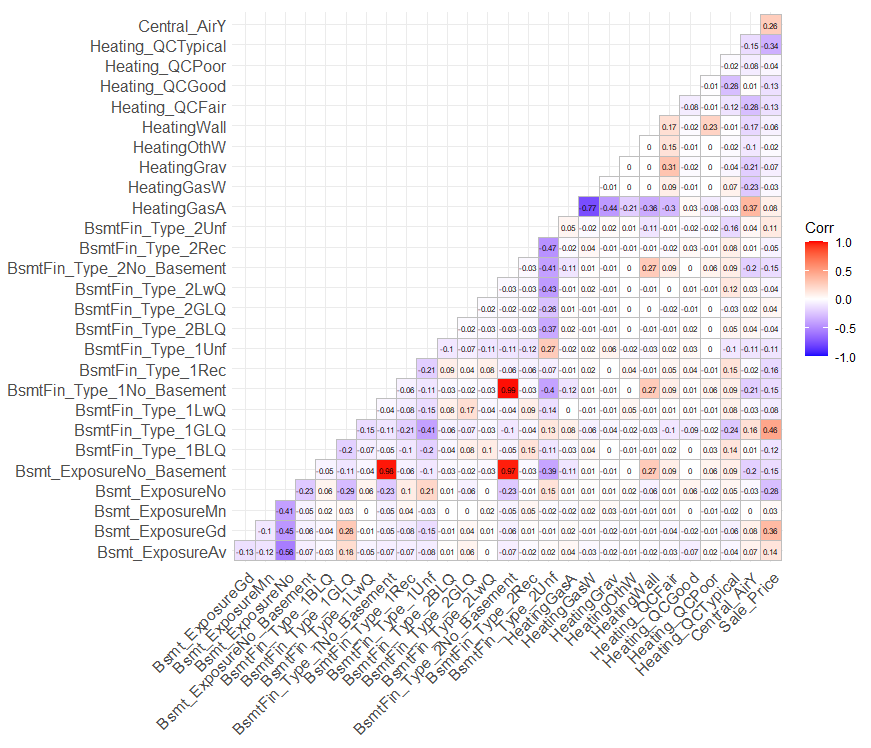
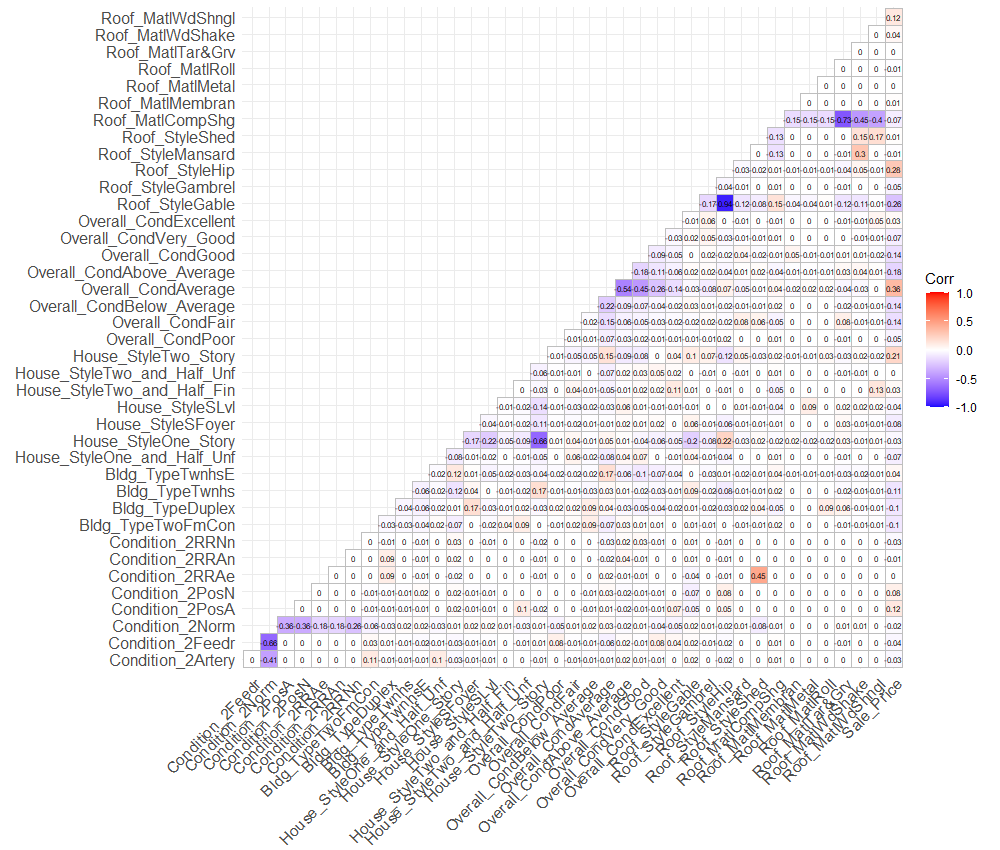
Nous savons aussi que le quartier a une importance majeure dans la détermination du prix. Cependant, la corrélation avec le prix est inférieure à 50% :

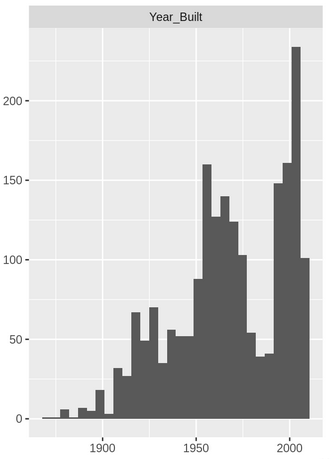
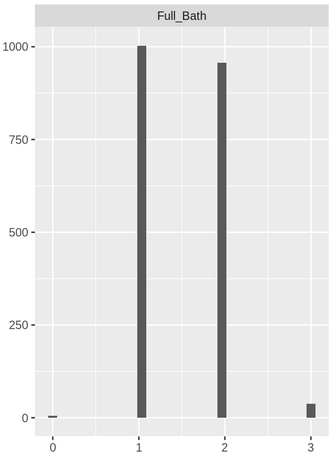
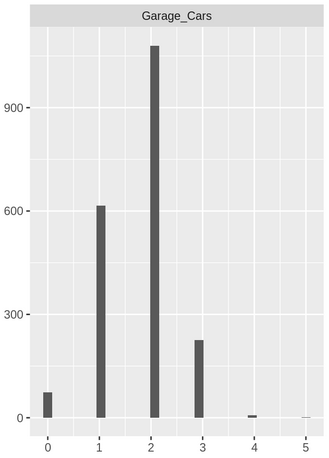
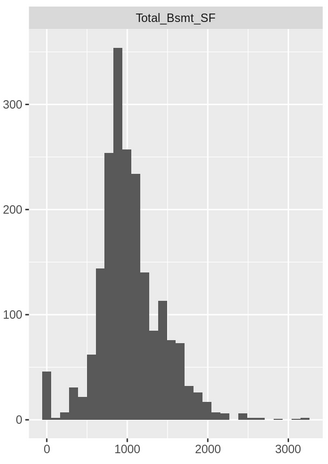
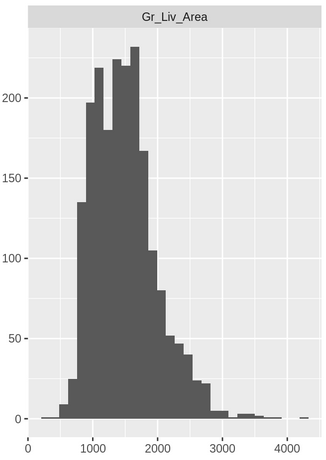
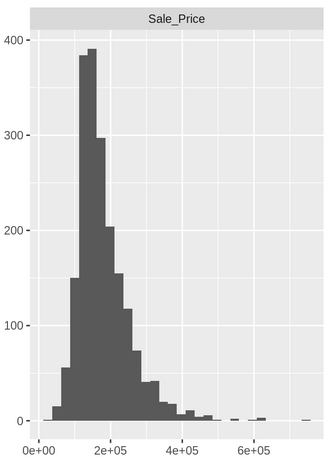
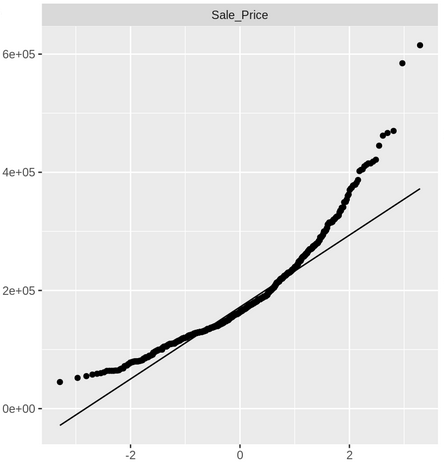
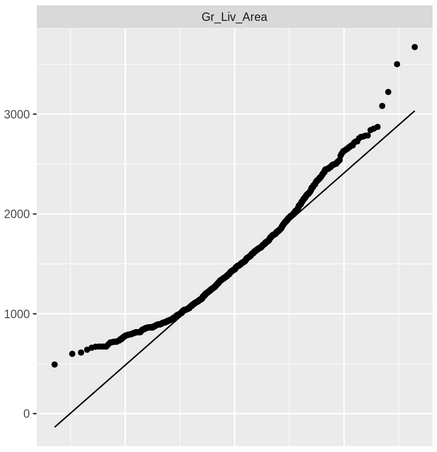
La plupart des maisons vendues sont à North Ames, mais les plus chères sont à Northridge Heights.

Malgré cette faible corrélation, nous sélectionnons tout de même cette variable, d’abord car le quartier est un élément intuitivement important, et aussi car nous verrons par la suite que cette variable est statistiquement importante dans nos modèles. En résumé :

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Variables corrélées sélectionnées | | | | | | | |
| Nom | **Définition** | | **Type** | | **Mesure** | | **Corrélation** |
| Sale\_Price | Prix de vente | | Int (continue) | | $ (USD) | | 100% |
| Gr\_Liv\_Area | Pieds carrés habitables au-dessus du rez-de-chaussé | | Int (continue) | | Pieds carrés (Square Feet) | | 71% |
| Garage\_Cars | Taille du garage en nombre de voitures | | Num (discrète) | | Nombre de voitures | | 65% |
| Total\_Basement\_SF | Pieds carrés totaux de la cave | | Num (continue) | | Pieds carrés (Square Feet) | | 63% |
| Year\_Built | Date initiale de construction | | Int (discrète) | | Année | | 56% |
| Full\_Bath | Salles de bains complètes au-dessus du rez-de-chaussé | | Int (discrète) | | Nombre de salle de bains | | 55% |
| Foundation | Type de fondation | | Factor (nominale) : 6 modalités | | Nom du matériau | | 52% (max, pour le béton coulé) |
| Neighborhood | Localisation physique au sein d’Ames | | Factor (nominale) : 29 modalités | | Nom de quartier | | 43% (max, pour Northridge Heights) |
| Variables corrélées non sélectionnées (multicolinéarité) | | | | | | | |
| Nom | | **Corrélation avec le prix de vente** | | **Variable explicative corrélée** | | **Corrélation avec cette variable** | |
| Garage\_Area | | 64% | | Garage\_Cars | | 89% | |
| TotRms\_AbvGrd | | 50% | | Gr\_Liv\_Area | | 81% | |
| Year\_Remod\_Add | | 53% | | Year\_Built | | 61% | |
| First\_Flr\_SF | | 62% | | Total\_Basement\_SF | | 80% | |

Voici d’autres graphiques non utilisés qui justifient que nous n’avons pas sélectionné d’autre variables, en raison d’une trop faible (< 50%) corrélation avec le prix de vente :





Analysons nos données graphiquement :

En effet, la surface habitable et le prix ont une asymétrie très semblable. Nous corrigerons par la suite l’asymétrie du prix de vente pour améliorer les analyses.

La surface habitable et le prix de vente ont l’air d’avoir une distribution similaire. Une analyse Quartile-Quartile devrait donner la même allure.

La plupart des maisons ont une surface de sous-sol d’environ 1 000 pieds carrés, avec beaucoup de maisons sans sous-sol. En effet, lorsque l’on a un sous-sol, il est rare qu’il soit extrêmement petit ou extrêmement grand.

La plupart des maisons ont 1 à 2 places pour voitures (2 étant le plus fréquent). Cela semble être un schéma assez courant dans les familles.

La plupart des maisons ont 1 à 2 baignoires (1 étant le plus fréquent). Ces valeurs correspondent bien à l’intuition.

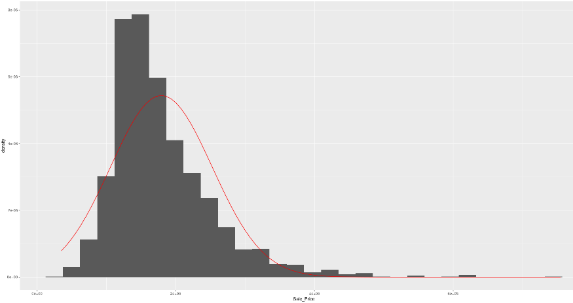
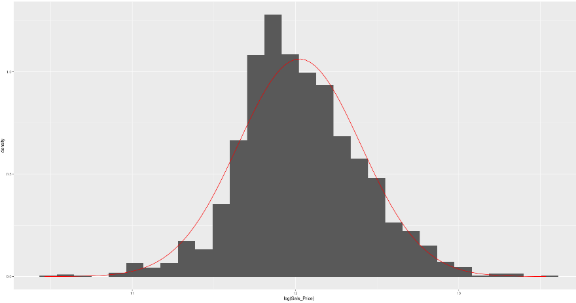
On remarque que malgré un grand nombre de maisons récentes, la plupart date d’avant 1975. On remarque surtout 2 moments : Le nombre de maisons construites après 1945 augmente drastiquement. Cela pourrait s’expliquer par la fin de la 2nd Guerre Mondiale ; Le nombre de maisons construites après 1970 diminue drastiquement. Cela pourrait s’expliquer par le fait que la crise pétrolière de 1973 ait eu un impact sur l’activité économique du pays.

Ainsi, une maison tirée au hasard possèdera très probablement les caractéristiques suivantes : un prix de 180 000$, une surface habitable au-dessus du rez-de-chaussée de 140 , 2 voitures, un sous-sol de 90 , 2 baignoires et construite en parpaings à North Ames en 1968.

Analysons nos données quantitativement :

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Statistiques | Fréquences | Distribution | Correctes | Manquantes |
| Sale\_Price [integer] | |  | | --- | | Mean (sd) : 179184.6 (73341.2) | | min ≤ med ≤ max: | | 35000 ≤ 161875 ≤ 755000 | | IQR (CV) : 82387.5 (0.4) | | 722 valeurs distinctes |  | 2002 (100.0%) | 0 (0.0%) |
| Gr\_Liv\_Area [integer] | |  | | --- | | Mean (sd) : 1494.3 (495.9) | | min ≤ med ≤ max: | | 334 ≤ 1445 ≤ 4316 | | IQR (CV) : 650.5 (0.3) | | 1085 valeurs distinctes |  | 2002 (100.0%) | 0 (0.0%) |
| Garage\_Cars [numeric] | |  | | --- | | Mean (sd) : 1.7 (0.7) | | min ≤ med ≤ max: | | 0 ≤ 2 ≤ 5 | | IQR (CV) : 1 (0.4) | | |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | 0 | : | 74 | ( | 3.7% | ) | | 1 | : | 615 | ( | 30.7% | ) | | 2 | : | 1080 | ( | 53.9% | ) | | 3 | : | 225 | ( | 11.2% | ) | | 4 | : | 7 | ( | 0.3% | ) | | 5 | : | 1 | ( | 0.0% | ) | |  | 2002 (100.0%) | 0 (0.0%) |
| Total\_Bsmt\_SF [numeric] | |  | | --- | | Mean (sd) : 1031.3 (401.4) | | min ≤ med ≤ max: | | 0 ≤ 974 ≤ 3206 | | IQR (CV) : 426.8 (0.4) | | 860 valeurs distinctes |  | 2002 (100.0%) | 0 (0.0%) |
| Year\_Built [integer] | |  | | --- | | Mean (sd) : 1967.5 (29.8) | | min ≤ med ≤ max: | | 1872 ≤ 1968 ≤ 2010 | | IQR (CV) : 46 (0) | | 113 valeurs distinctes |  | 2002 (100.0%) | 0 (0.0%) |
| Full\_Bath [integer] | |  | | --- | | Mean (sd) : 1.5 (0.5) | | min ≤ med ≤ max: | | 0 ≤ 1 ≤ 3 | | IQR (CV) : 1 (0.4) | | |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | 0 | : | 5 | ( | 0.2% | ) | | 1 | : | 1003 | ( | 50.1% | ) | | 2 | : | 957 | ( | 47.8% | ) | | 3 | : | 37 | ( | 1.8% | ) | |  | 2002 (100.0%) | 0 (0.0%) |
| Foundation [factor] | |  | | --- | | 1.BrkTil : Brick & Tile :Brique & Tuile | | 2.CBlock : Cinder Block : Parpaing | | 3. PConc : Poured Concrete : Béton Coulé | | 4. Slab : Dalle | | 5. Stone : Pierre | | 6. Wood : Bois | | |  |  |  |  | | --- | --- | --- | --- | | 248 | ( | 12.4% | ) | | 903 | ( | 45.1% | ) | | 812 | ( | 40.6% | ) | | 27 | ( | 1.3% | ) | | 8 | ( | 0.4% | ) | | 4 | ( | 0.2% | ) | |  | 2002 (100.0%) | 0 (0.0%) |
| Neighborhood [factor] | |  | | --- | | 1. North\_Ames | | 2. College\_Creek | | 3. Old\_Town | | 4. Edwards | | 5. Somerset | | 6.Northridge\_Heights | | 7. Gilbert | | 8. Sawyer | | 9. Northwest\_Ames | | 10. Sawyer\_West | | [ 19 others ] | | |  |  |  |  | | --- | --- | --- | --- | | 360 | ( | 18.0% | ) | | 213 | ( | 10.6% | ) | | 177 | ( | 8.8% | ) | | 129 | ( | 6.4% | ) | | 67 | ( | 3.3% | ) | | 67 | ( | 3.3% | ) | | 128 | ( | 6.4% | ) | | 121 | ( | 6.0% | ) | | 113 | ( | 5.6% | ) | | 89 | ( | 4.4% | ) | | 538 | ( | 26.9% | ) | |  | 2002 (100.0%) | 0 (0.0%) |

**3. Modèle**

Tout d’abord, nos modèles seront de la forme log-niveau. En effet la distribution du Sale\_Price est assez asymétrique, nous allons prendre son logarithme ce qui va permettre de réduire l’asymétrie, stabiliser la variance et donc améliorer les modèles. Le Sale\_Price (à gauche) semble plutôt suivre une loi Gamma. Une fois en log (à droite), il semble suivre une loi normale. Effectivement, selon le test d’Anderson-Darling, il est très probable (α = 5%) que le log(Sale\_Price) suive une loi normale (p-value > 0.05, donc on ne rejette pas l’hypothèse nulle).

Nous aurrons en tout 4 modèles : les 3 premiers sont spécifiés manuellement et le dernier est un compromis entre le ajusté, le Cp de Mallow et le critère d’information Bayésien (BIC), uniquement pour les variables quantitatives et pour chacun des modèles possibles (.

1er modèle :

Ce modèle n’utilise que GrLivArea pour expliquer le prix de vente, ceci est fait pour vérifier si le GrLivArea peut expliquer à lui seul le prix d’un bien, puisque sa corrélation avec le prix est de loin la plus élevée.

2ème modèle :

Ce modèle reprend les variables sélectionnées précédemment (2.2) selon leur corrélation.

3ème modèle :

Ce modèle utilise toutes les variables pour expliquer le prix. Son ajusté permet d’être comparé avec les autres modèles pour vérifier si finallement toutes les variables ne sont pas importantes. Les coefficients ne seront pas précisés en raison du nombre trop conséquent de variables, mais seulement le ajusté.

4ème modèle :

Ces variables ont été choisi de la manière suivante :

1. Test de tous les modèles possibles pour les variables quantitatives (donc modèles à tester, soit modèles) avec un algorithme Branch and Bound.
2. Pour chacun de ces modèles, extraire son ajusté, son Cp et son BIC
3. Trouver le modèle qui fait le compromis entre un ajusté élevé et un Cp et BIC faible
4. Utiliser les variables de ce modèle

Voici les résultats avec en bleu les meilleurs modèles selon chaque critères et en orange le choix du modèle qui est le meilleur compromis entre ces 3 critères.

ADJR2 CP BIC

Il est important de noter que nous travaillons ici uniquement avec une base de données constituée des variables quantitative de la base de données initiale. Aussi, bien que nous cherchons le meilleur modèle selon ces 3 critères, nous savons déjà qu’il nous est proposé un modèle avec des variables corrélées entre elles (GarageCars et GarageArea) comme nous l’avions vu dans l’analyse de corrélation précédente. Nous testerons plus tard cette hypothèse avec l’indicateur VIF.

15 0.8825505 52.19139 -4181.210

16 0.8830563 44.46818 -4183.258

17 0.8835151 37.56253 -4184.535

18 0.8838898 32.11240 -4184.391

19 0.8841371 28.85509 -4182.069

20 0.8843700 25.85103 -4179.506

21 0.8845433 23.87344 -4175.918

22 0.8846395 23.22205 -4170.996

23 0.8847232 22.78705 -4165.858

24 0.8847981 22.50345 -4160.571

25 0.8848568 22.49961 -4155.001

26 0.8848932 22.87918 -4149.045

27 0.8849114 23.56891 -4142.774

28 0.8849131 24.54286 -4136.216

29 0.8848726 26.23765 -4128.924

30 0.8848257 28.04010 -4121.523

31 0.8847692 30.00715 -4113.955

32 0.8847109 32.00290 -4106.357

33 0.8846525 34.00000 -4098.758

Nous choisissons le modèle 21, dont les variables sont explicitées dans le modèle ci-dessus.

**4. Résultats**

4.1. Statistiques descriptives

On remarque que la plupart des maisons sont vendues entre 350k et 500k $, et qu’elles ont comme fondation du béton.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Tableau 1 : Types de fondation(nombre) selon le prix (en $) | | | | | | |
|  | BrkTil | CBlock | PConc | Slab | Stone | Wood |
| (350k,500k] | 2 | 5 | 45 | 0 | 0 | 0 |
| (500k,600k] | 0 | 0 | 3 | 0 | 0 | 0 |
| (600k,700k | 0 | 0 | 3 | 0 | 0 | 0 |
| (700k,755k] | 0 | 0 | 1 | 0 | 0 | 0 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | BrkTil | CBlock | PConc | Slab | Stone | Wood |
| North\_Ames | 4 | 331 | 17 | 7 | 1 | 0 |
| College\_Creek | 1 | 36 | 176 | 0 | 0 | 0 |
| Old\_Town | 86 | 53 | 32 | 2 | 4 | 0 |
| Edwards | 18 | 78 | 22 | 11 | 0 | 0 |
| Somerset | 0 | 0 | 67 | 0 | 0 | 0 |
| Northridge\_Heights | 0 | 0 | 67 | 0 | 0 | 0 |
| Gilbert | 0 | 2 | 124 | 1 | 0 | 1 |
| Sawyer | 5 | 104 | 9 | 3 | 0 | 0 |
| Northwest\_Ames | 0 | 92 | 21 | 0 | 0 | 0 |
| Sawyer\_West | 2 | 17 | 70 | 0 | 0 | 0 |
| Mitchell | 0 | 47 | 35 | 1 | 0 | 1 |
| Brookside | 53 | 28 | 13 | 2 | 0 | 0 |
| Crawford | 31 | 29 | 17 | 0 | 1 | 0 |
| Iowa\_DOT\_and\_Rail\_Road | 28 | 23 | 9 | 0 | 2 | 0 |
| Timberland | 0 | 15 | 32 | 0 | 0 | 2 |
| Northridge | 0 | 1 | 66 | 0 | 0 | 0 |
| Stone\_Brook | 0 | 0 | 13 | 0 | 0 | 0 |
| South\_and\_West\_of\_Iowa\_State\_University | 18 | 7 | 9 | 0 | 0 | 0 |
| Clear\_Creek | 2 | 26 | 9 | 0 | 0 | 0 |
| Meadow\_Village | 0 | 0 | 0 | 0 | 0 | 0 |
| Briardale | 0 | 0 | 0 | 0 | 0 | 0 |
| Bloomington\_Heights | 0 | 0 | 1 | 0 | 0 | 0 |
| Veenker | 0 | 14 | 3 | 0 | 0 | 0 |
| Northpark\_Villa | 0 | 0 | 0 | 0 | 0 | 0 |
| Blueste | 0 | 0 | 0 | 0 | 0 | 0 |
| Greens | 0 | 0 | 0 | 0 | 0 | 0 |
| Green\_Hills | 0 | 0 | 0 | 0 | 0 | 0 |
| Landmark | 0 | 0 | 0 | 0 | 0 | 0 |
| Hayden\_Lake | 0 | 0 | 0 | 0 | 0 | 0 |

Les maisons les plus vendues à North Ames sont en parpaings.

Tableau 2 : Types de fondation(nombre) selon le quartier

On remarque que la plupart des maisons sont vendues vers Northridge Heights entre 350k et 500k $.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Tableau 3: Nombre de quartier par classe de prix | | | | |
|  | (350k,500k] | (500k,600k] | (600k,700k] | (700k,750k] |
| North\_Ames | 0 | 0 | 0 | 0 |
| College\_Creek | 0 | 0 | 0 | 0 |
| Old\_Town | 1 | 0 | 0 | 0 |
| Edwards | 1 | 0 | 0 | 0 |
| Somerset | 1 | 0 | 0 | 0 |
| Northridge\_Heights | 24 | 1 | 2 | 0 |
| Gilbert | 1 | 0 | 0 | 0 |
| Sawyer | 0 | 0 | 0 | 0 |
| Northwest\_Ames | 0 | 0 | 0 | 0 |
| Sawyer\_West | 0 | 0 | 0 | 0 |
| Mitchell | 0 | 0 | 0 | 0 |
| Brookside | 0 | 0 | 0 | 0 |
| Crawford | 1 | 0 | 0 | 0 |
| Iowa\_DOT\_and\_Rail\_Road | 0 | 0 | 0 | 0 |
| Timberland | 5 | 0 | 0 | 0 |
| Northridge | 11 | 1 | 1 | 1 |
| Stone\_Brook | 5 | 1 | 0 | 0 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Tableau 4 : Résultats des estimations MCO sur le salaire | | | | |
| Variable expliquée : log(SalePrice) | Modèle 1 | Modèle 2 | Modèle 3 | Modèle 4 |
| Variables explicatives : |  |  |  |  |
| GrLivArea | 0.0005\*\*\* | 0.0003\*\*\* |  |  |
| (0.00001) | (0.00001) |  |  |
| GarageCars |  | 0.07\*\*\*  (0.007) |  |  |
| TotalBsmtSF |  | 0.0001\*\*\* |  |  |
|  | (0.00001) |  |  |
| YearBuilt |  | 0.002\*\*\* |  | 0.002\*\*\* |
|  | (0.0003) |  | (0.0001) |
| FullBath |  | 0.001  (0.01) |  |  |
| FoundationCBlock (ref:BrkTill) |  | 0.03\*  (0.01) |  |  |
| FoundationPConc |  | 0.04\*  (0.01) |  |  |
| FoundationSlab |  | -0.01  (0.04) |  |  |
| FoundationStone |  | 0.05  (0.09) |  |  |
| FoundationWood |  | -0.06\*  (0.03) |  |  |
| NeighborhoodCollegeCreek  (ref:North Ames ) |  | 0.02  (0.01) |  |  |
| NeighborhoodOld\_Town |  | -0.04\*  (0.01) |  |  |
| …(17 levels ignorés, trop long) |  | …  (…) |  |  |
| NeighborhoodVeenker |  | 0.1\*\*  (0.04) |  |  |
| Lot\_Frontage |  |  |  | 0.0002\*  (0.00008) |
| Lot\_Area |  |  |  | 0.000002\*\*\*  (0.0000004) |
| Year\_Remod\_Add |  |  |  | 0.002\*\*\*  (0.0002) |
| Mas\_Vnr\_Area |  |  |  | 0.00005\*\*  (0.00001) |
| BsmtFin\_SF\_1 |  |  |  | -0.005\*\*  (0.001) |
| BsmtFin\_SF\_2 |  |  |  | -0.00005\*\*  (0.00001) |
| Bsmt\_Unf\_SF |  |  |  | -0.00005\*\*\*  (0.00001) |
| Total\_Bsmt\_SF |  |  |  | 0.0002\*\*\*  (0.00001) |
| First\_Flr\_SF |  |  |  | 0.0002\*\*\*  (0.00002) |
| Second\_Flr\_SF |  |  |  | 0.0003\*\*\*  (0.00001) |
| Bsmt\_Full\_Bath |  |  |  | 0.01\*  (0.007) |
| Bedroom\_AbvGr |  |  |  | -0.02\*\*  (0.006) |
| Kitchen\_AbvGr |  |  |  | -0.2\*\*\*  (0.04) |
| TotRms\_AbvGrd |  |  |  | 0.01\*\*  (0.004) |
| Fireplaces |  |  |  | 0.04\*\*\*  (0.005) |
| Garage\_Cars |  |  |  | 0.04\*\*\*  (0.008) |
| Garage\_Area |  |  |  | 0.0001\*\*\*  (0.00003) |
| Wood\_Deck\_SF |  |  |  | 0.00008\*\*\*  (0.00002) |
| Screen\_Porch |  |  |  | 0.0002\*\*\*  (0.00005) |
| Pool\_Area |  |  |  | -0.0001  (0.0001) |
|  |  |  |  |  |
| Constante | 11.14 | 5.49 | -73.14 | 1.18 |
| (0.018) | (185,8) | (41.73) | (0.3) |
| N | 2002 | 2002 | 2002 | 2002 |
| R² ajusté | 0.59 | 0.86 | 0.94 | 0.88 |
| Test de Fisher | Rejet H0 | Rejet H0 | Rejet H0 | Rejet H0 |
| Test d’hétéroscédasticité | Rejet H0 | Rejet H0 | Rejet H0 | Rejet H0 |
| Correction de l’hétéroscédasticité | Oui | Oui | Oui | Oui |
| Normalité des résidus | Rejet H0 | Rejet H0 | Rejet H0 | Rejet H0 |
| Multi-colinéarité | N/A | Aucune | N/A | GarageCars |
| Autocorrélation | Rejet H0 | Rejet H0 | Rejet H0 | Rejet H0 |
| Linéarité | Non Rejet H0 | Non Rejet H0 | Rejet H0 | Non Rejet H0 |

Les écarts types sont entre parenthèses. \* *p* < 0.05, \*\* *p* < 0.01, \*\*\* *p* < 0.001

On remarque d’abord des ajustés élevés, des statistiques de Fisher significatives, beaucoup d’hétéroscédasticité, de non-normalité et d’autocorrélation des résidus. Ces 4 modèles ont été corrigés de l’hétéroscédasticité par la méthode de White. Les estimations non-corrigées sont cependant consultables via ce [lien](https://gist.githubusercontent.com/gregory798/b863f8aecb968a234a927d27b426f7db/raw/f59e0fe988cc8a33297d0486c23506627d43d2c1/MCO%2520non%2520corrig%25C3%25A9s.txt). Dans le modèle 1, nous avons volontairement omis 17 modalités de la variable Neighborhood (le quartier) à des fins de lisibilité. Dans le modèle 3, nous n’avons pas affiché les estimations car c’est un modèle avec toutes les variables, soit 71 variables, dont certaines variables catégorielles se divisent elles aussi en dizaines de modalités, ce qui rend le rapport trop long. Cependant, ces estimations sont tout de même consultables via ce [lien](https://gist.githubusercontent.com/gregory798/2060fab1618dc93e2efbb779a5bb2d75/raw/870bcfc9823fe094f3da41d9a7fb77be16328216/Mod%25C3%25A8le%25203%2520corrig%25C3%25A9.txt). Une interprétation du modèle 1 corrigé serait de dire que lorsque le nombre de pied carrés au-dessus du rez-de-chaussée augmente d’une unité, le prix augmente en moyenne de 0.05%. Lorsque la surface au-dessus du rez-de-chaussée est nulle, le prix de vente est en moyenne de soit environ 70k$. Dans le modèle 4, on remarque que, toutes choses égales par ailleurs, lorsque l’on rajoute une chambre au-dessus du rez-de-chaussée, le prix baisse en moyenne de 2%, ce qui, à priori, n’est pas très intuitif. Cependant, selon le recensement américain [3], on sait que les foyers américains comptent de moins en moins de personnes, et donc que les biens les plus recherchés sont ceux avec peu de chambres. De manière plus logique, on comprend que rajouter une chambre sans augmenter la surface totale revient à diminuer ou diviser des espaces, ce qui peut rendre la maison moins attractive. Personne ne veut d’une maison avec un nombre trop élevé de chambres.

**5. Conclusion**

Pour résumer, nous avons analysé les différents facteurs qui rentrent en jeux en ce qui concerne la détermination du prix de vente d’une maison. Nous avons conclu que cette base de données n’est pas faite pour des régressions linéaires, bien que nous le supposions auparavant. En effet, on remarque que peu d’hypothèses sont vérifiées, et donc beaucoup que d’hypothèse importantes ne sont pas vérifiées, comme l’homoscédasticité, la distribution normale des résidus ou encore l’autocorrélation. Ceci pourrait s’expliquer par l’existence de clusters, notamment les quartiers, puisque le prix d’une maison peut être proche de la moyenne de celui de son quartier, mais très éloignée d’un autre quartier. Malgré cela, le ajusté reste très élevé globalement. Pour spécifier un modèle meilleur, il faudrait partir de celui avec toutes les variables, enlever chaque variable unes par unes, puis vérifier s’il y a des différences statistiquement significatives. En effet, avec « seulement » 20 variables, nous arrivons à un ajusté élevé, 0.88 (contre 0.94 mais 71 variables pour le modèle 3) mais en ayant aussi un Cp de Mallow assez bas (23.87), ainsi qu’un critère d’information bayésien bas (- 4 175).

Cette base de données reste assez particulière au niveau des corrélations, puisque beaucoup de variables sont corrélées entre elles, même lorsque les coefficients de corrélation ne sont pas très élevés. En effet, l’étude portant sur des biens immobiliers, on s’attend logiquement à des liens inter-facteurs et des liens inter-clusters. Inter-facteurs dans le sens où, lorsqu’une maison est plus grande que la moyenne, il est plus probable que la surface de son garage soit aussi plus grande, par exemple. Or ce lien est poussé à l’extrême dans des cas des maisons puisqu’une plus grande surface signifie automatiquement plus de chambres etc. Inter-clusters dans le sens où, les maisons d’un même quartier ont de très fortes chances de se ressembler dans les caractéristiques, que ce soit dans le type de fondation, la surface, le garage etc. Cela implique donc que les estimations par MCO ne sont pas fiables, notamment les t-values en raison d’écarts-types peu fiables. A noter que toutes ces analyses ne concernent que la ville d’Ames dans l’Iowa, et donc qu’aucune conclusion à l’échelle globale ne peut être tirée.

Il pourrait être intéressant de faire les mêmes analyses mais avec des maisons dispersées selon la commune, ou à toute autre échelle plus grande, afin de réduire l’effet des quartiers sur la variance et rapprocher les prix de vente entre communes du prix de vente moyen global.

**Bibliographie**

[1] De Cock, D., 2011. Ames, Iowa: Alternative to the Boston housing data as an end of semester regression project. Journal of Statistics Education, 19(3). [[lien](http://jse.amstat.org/v19n3/decock/DataDocumentation.txt)]

[2] Bureau of Labor Statistics, U.S. Department of Labor. Consumer Expenditures – 2020 [[lien](https://www.bls.gov/news.release/pdf/cesan.pdf)]

[3] Current Population Survey, 2020 ASEC Technical Documentation [[lien](https://www2.census.gov/programs-surveys/cps/techdocs/cpsmar20.pdf)]

**Annexe**

|  |  |  |  |
| --- | --- | --- | --- |
| Modèle | Test | P-value | H0 |
| 1,2,3,4 | Fisher | < 0.05 | Rejet H0 |
| 1,2,3,4 | Non-Constant Error Variance | < 0.05 | Rejet H0 |
| 1,2,3,4 | Breusch-Pagan | < 0.05 | Rejet H0 |
| 1,2,3,4 | Shapiro-Wilk | < 0.05 | Rejet H0 |
| 1,3 | VIF | N/A | N/A |
| 2 | VIF | Aucune variable | N/A |
| 4 | VIF | Une (1) variable | GarageCars |
| 1,2,3,4 | Durbin-Watson | < 0.05 | Rejet H0 |
| 1,2,4 | Rainbow | > 0.05 | Non Rejet H0 |
| 3 | Rainbow | < 0.05 | Rejet H0 |
| 1,2,3,4 | Distance de Cook | Aucune variable influente | N/A |
| Sale\_Price | Anderson-Darling | < 0.05 | Rejet H0 |
| log(Sale\_Price) | Anderson-Darling | > 0.05 | Non Rejet H0 |