

UNIVERSITÉ DE PARIS
FACULTÉ DE DROIT, ECONOMIE ET GESTION

Facteurs explicatifs du prix de vente d'une maison.

Prix et caractéristiques des maisons à Ames, Iowa (USA).

Grégory BOURNASSENKO

Romain WU

Axel ATHOR

Groupe de TD : N° 4

Chargé de TD : M. Dondjio

LICENCE 3^{ÈME} 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, aux États-Unis d'Amérique, ainsi qu'un certain nombre de variables intuitivement explicatives (par exemple la surface habitable, le nombre de chambres, etc.) que l'on pourrait appeler « caractéristiques ». Nous nous intéressons donc aux facteurs qui peuvent expliquer le prix d'un bien immobilier, bien que l'analyse porte uniquement sur cette ville des US, et donc que les conclusions de cette analyse seront vraies surtout pour cette ville aux États-Unis, en particulier.

Nous partons du principe que chaque maison possède des points semblables, mais aussi des différences et ce 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, l'isolement par rapport aux autres quartiers, etc.). 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 qui influencent le 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. Ces prises de décision peuvent donc être accélérées et/ou améliorées grâce à la connaissance des variables les plus explicatives dans la détermination du prix, car elles peuvent servir d'heuristique [3]. Par exemple, à prix égal, un ménage préférera un bien pour lequel une variable explicative est plus présente plutôt que pour un autre bien. Cette prise de décision peut se faire de la manière suivante : lorsque l'on souhaite acheter le meilleur article d'une catégorie, une heuristique courante serait de dire que les meilleurs articles sont les plus chers. Ce n'est pas toujours vrai, mais cela permet de simplifier et d'accélérer notre prise de décision. Ainsi, on pourrait se dire que les meilleures maisons sont celles dont la surface habitable est la plus élevée (par exemple, si telle est la conclusion.).

Nous nous poserons donc la question suivante : Quels sont les facteurs qui influencent le plus significativement 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 facilement accessible de plusieurs manières, par exemple :

- Depuis le package “[modeldata](#)”, en important la base de données “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 compilée par [Dean De Cock](#) pour être utilisée dans le cadre d’un enseignement des sciences des données. Il s’agit d’une alternative pour les spécialistes des données comme les étudiants qui recherchent une version modernisée et étendue de la base de données sur les logements de Boston, souvent citée.

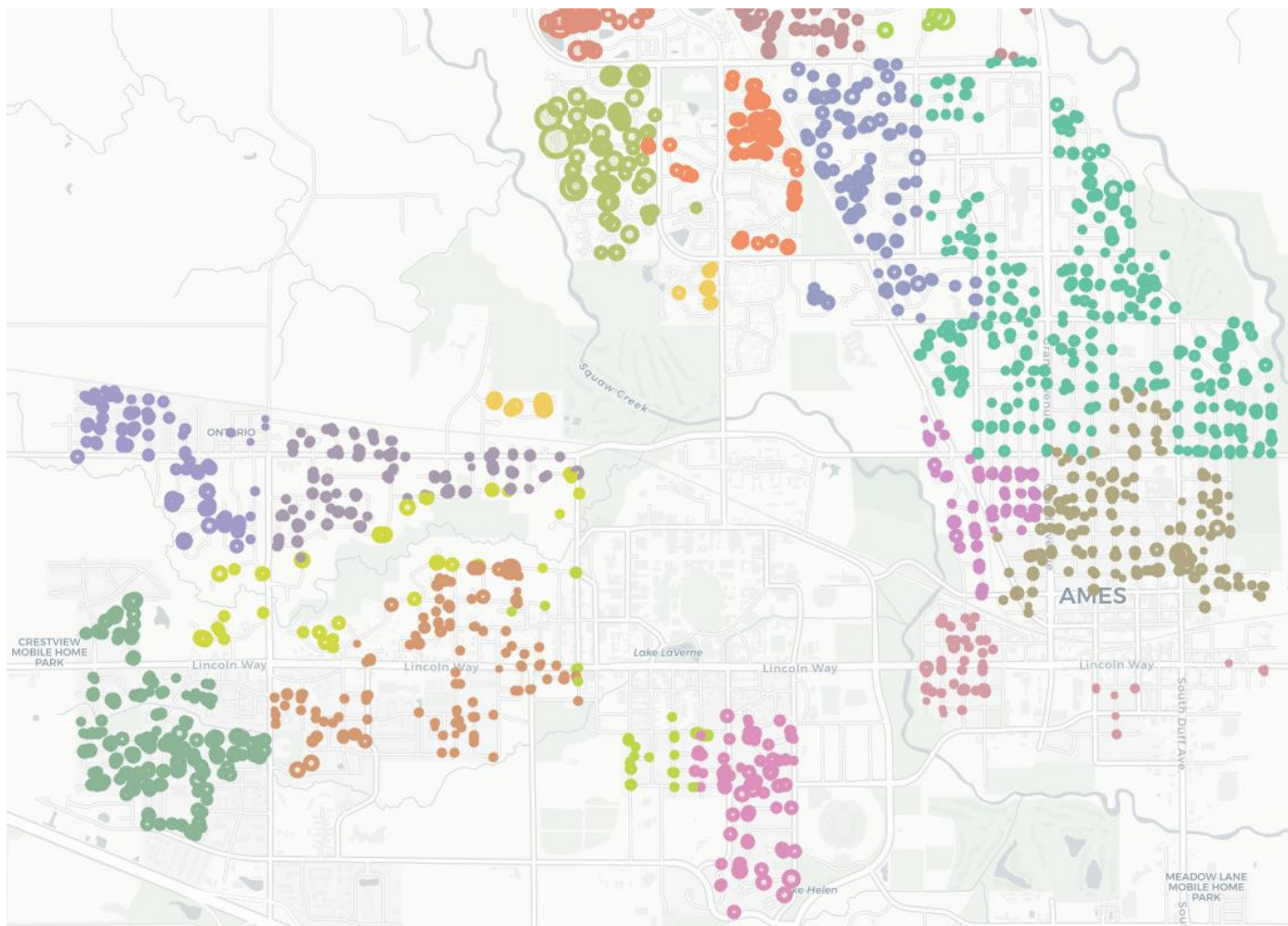
Il s’agit donc d’un grand échantillon de données relatives aux maisons, dispersées dans plusieurs quartiers de la ville d’Ames, collectées de 2006 à 2010. Les dimensions (initiales) de cet échantillon sont de 2930 lignes/observations pour 74 colonnes/variables. Attention à ne pas confondre Ames dans l’Iowa (66 000 habitants) et Ames dans le Pas-de-Calais (652 habitants).

Avec 74 variables, nous allons devoir nous séparer de certaines, pour nous intéresser aux plus pertinentes. Cette pertinence sera définie de plusieurs manières au fil de la recherche. La définition précise de chaque variable (en anglais) est disponible via ce [lien](#) et en français [ici](#).

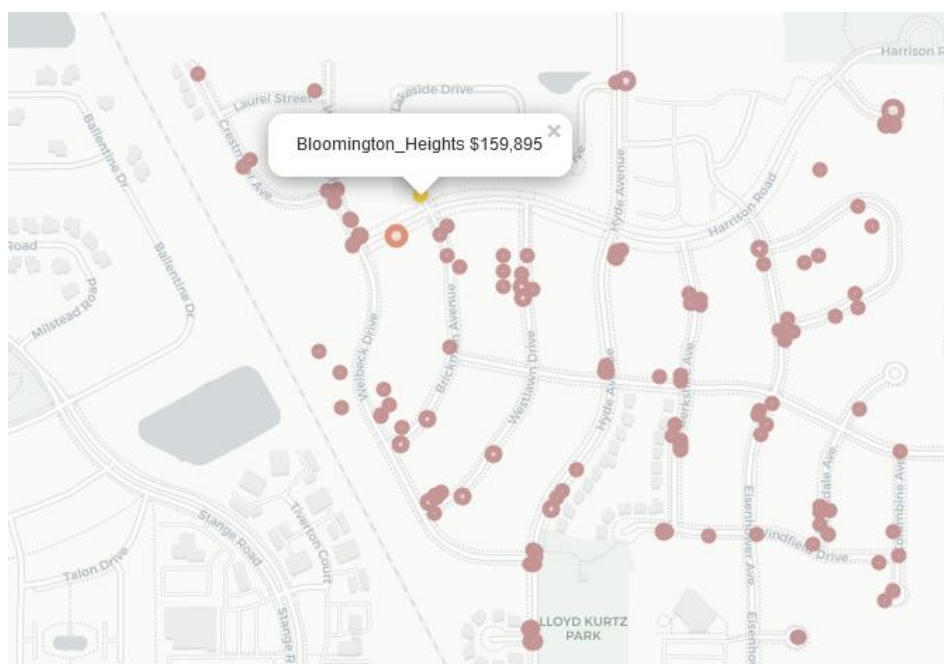
Cette base de données ne contient initialement aucune donnée manquante. Elle est constituée (initialement) de 23 variables nominales, 23 ordinales, 14 discrètes et 20 continues.

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

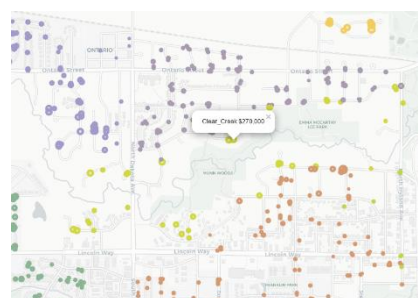
Par souci de simplicité, et pour rendre l’analyse la plus universelle possible, c’est-à-dire en écartant les situations qui seraient trop inhérentes et spécifiques à la ville d’Ames et/ou certains types de vente ou d’habitation, nous travaillerons uniquement avec les ventes normales et les habitations à une (1) famille. Une fois ce nettoyage fait, nous nous retrouvons avec une base de données de dimensions 2 002 x 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) :



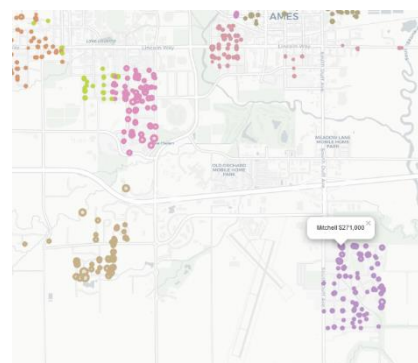
Il n'y a pas de données au centre de la ville car c'est l'emplacement d'une université.



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



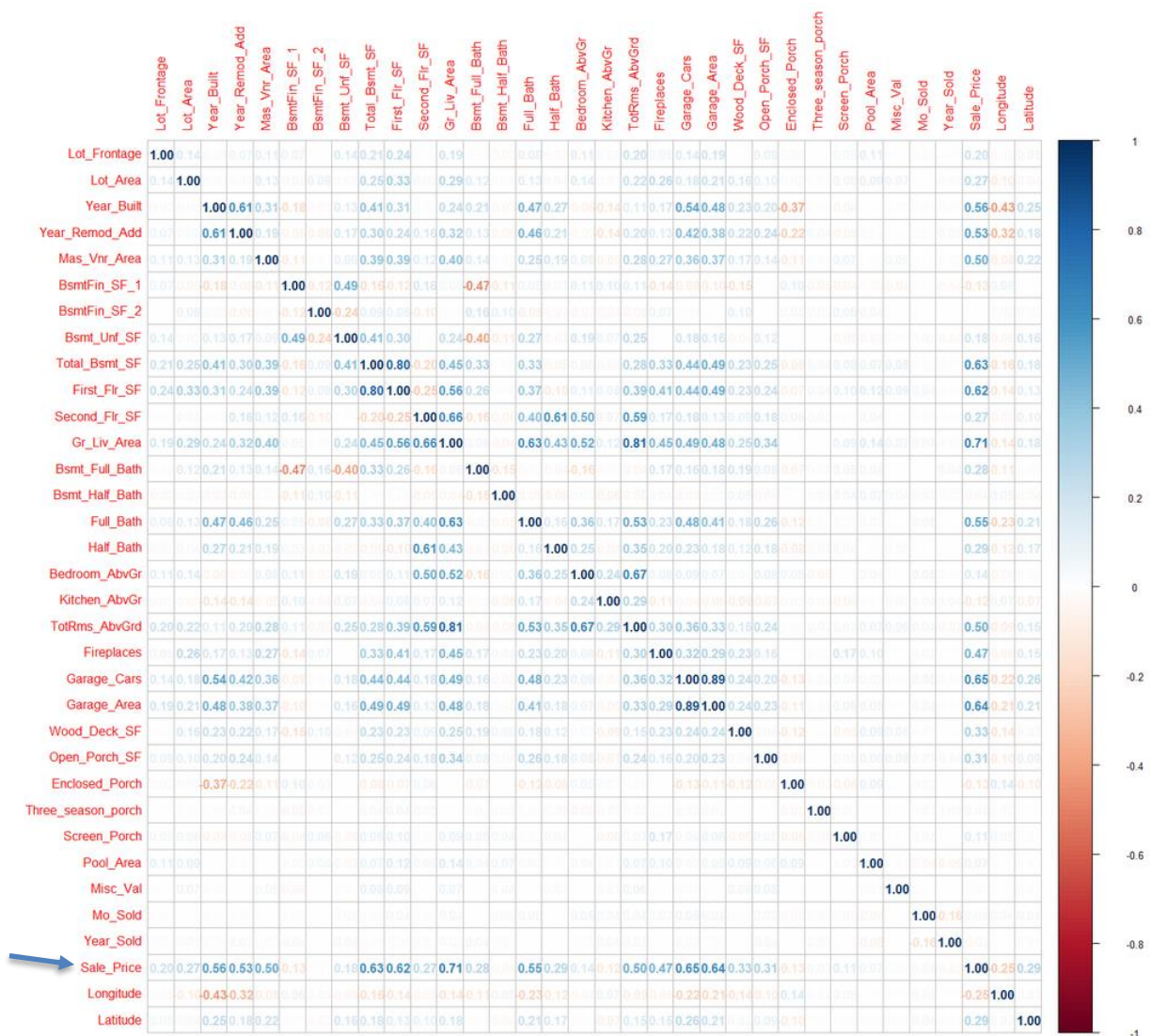
Il y a des quartiers qui sont mal délimités.



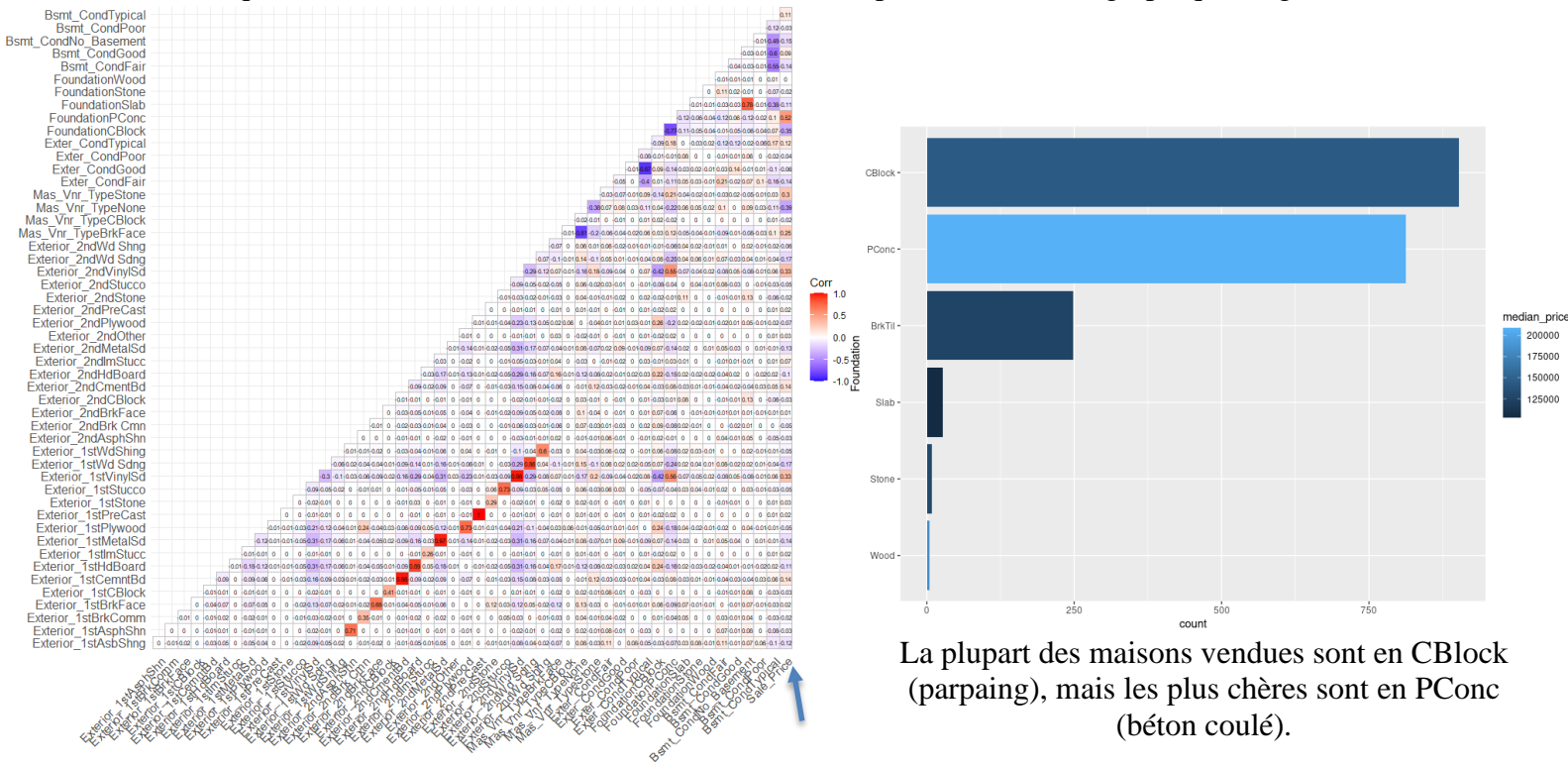
Il y a des quartiers assez isolés des autres.

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» (qui devient alors toujours égale à la modalité « 1Fam ») et « Sale Condition» (qui devient alors toujours égale à la modalité « Normal »), puisqu'elles sont maintenant constantes et donc que la variance est nulle pour ces variables. Nous allons donc devoir les supprimer (et cela n'affecte en rien les futurs modèles, puisqu'aucune estimation ne peut se faire avec des variables de ce type.). 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, le R^2 ajusté vaut 0.9463, dans ce modèle, donc la marge de progression pour extraire du terme d'erreur des variables qui pourraient mieux expliquer $\log(\text{Sale_Price})$ 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 :



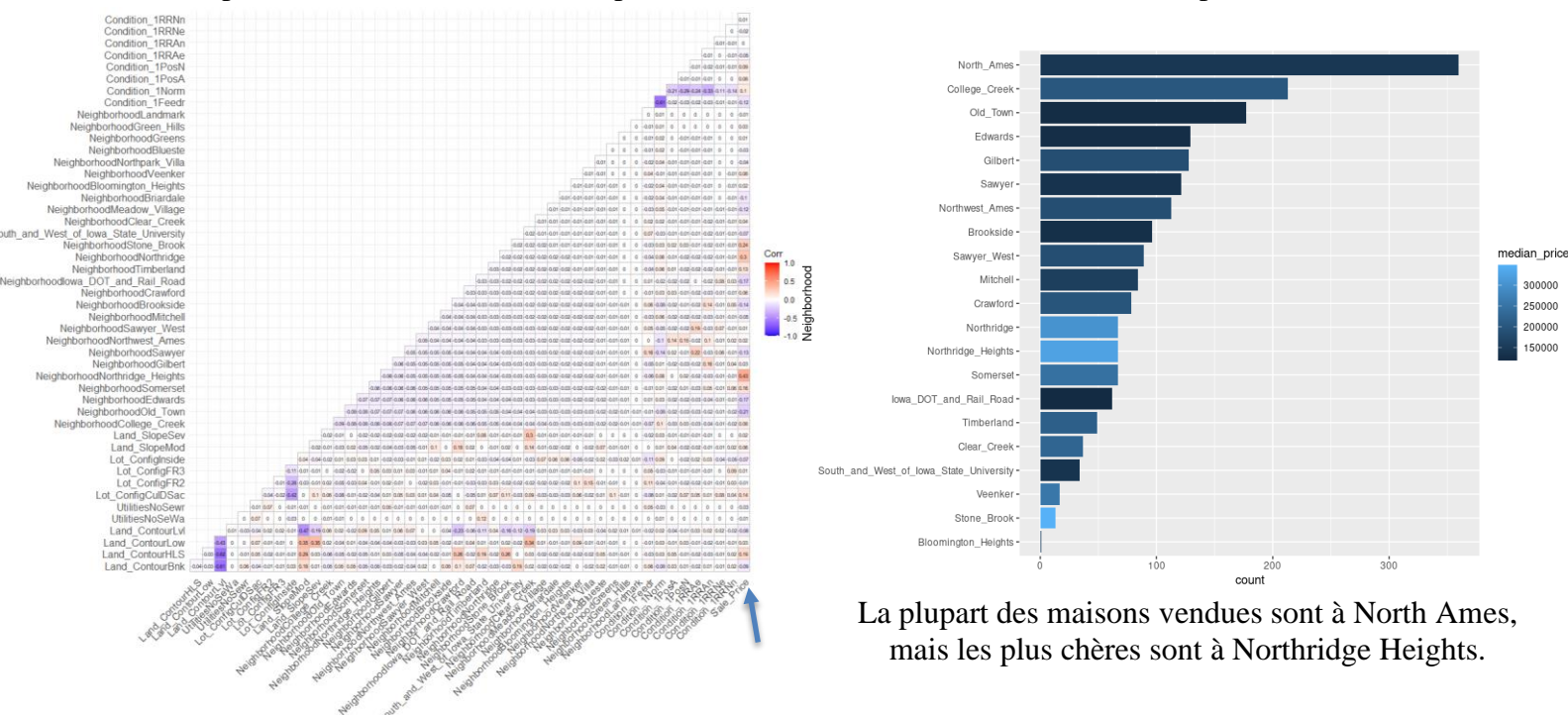
On remarque qu'il y a une corrélation supérieure à 50% avec le type de fondation, notamment lorsque les fondations sont en béton coulé, comme on peut le voir sur le graphique de gauche :



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, sa corrélation avec le prix de vente est inférieure à 50%, comme on peut le voir :

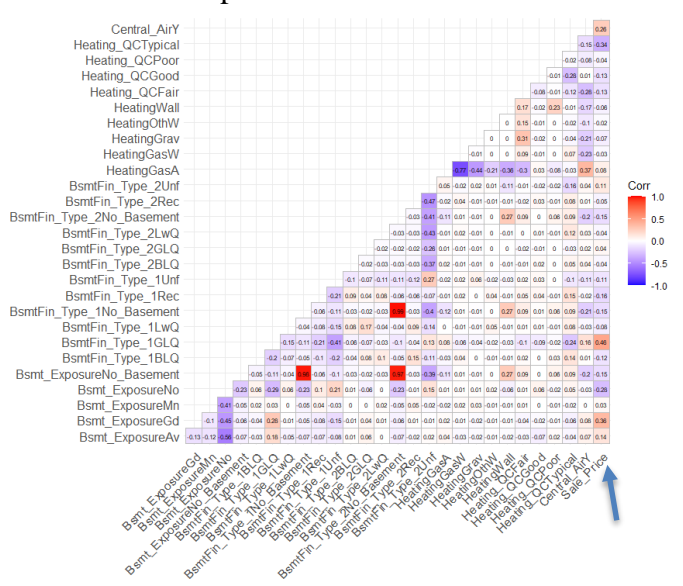
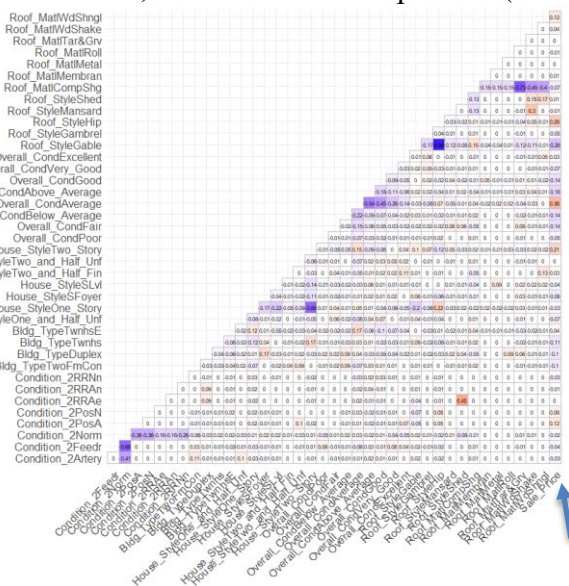


La plupart des maisons vendues sont à North Ames, mais les plus chères sont à Northridge Heights.

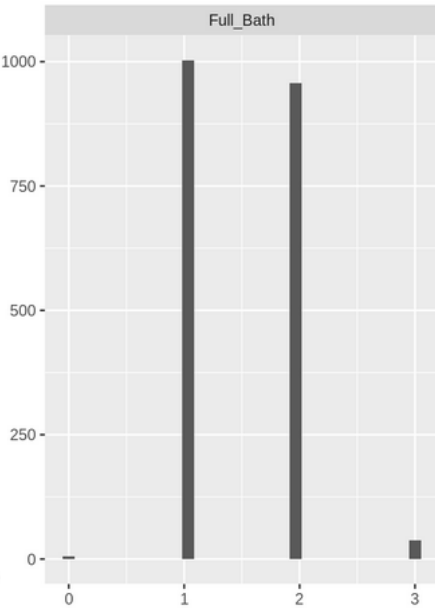
Malgré cette faible corrélation, nous sélectionnons tout de même cette variable, d'abord, car le quartier est un élément intuitivement important, et aussi, car nous verrons par la suite que cette variable est statistiquement importante (« significative ») dans nos modèles. En résumé :

Variables corrélées sélectionnées				
Nom de la variable	Définition	Type	Mesure	Corrélation
Sale_Price	Prix de vente	Int (continue)	\$ (USD)	100%
Gr_Liv_Area	Surface/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	Surface/Pieds carrés totaux du sous-sol	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	Nom du quartier d'appartenance au sein d'Ames	Factor (nominale) : 29 modalités	Nom du quartier	43% (max, pour Northridge Heights)
Variables corrélées <u>non</u> sélectionnées (pour éviter toute multi colinéarité)				
Nom de la variable	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%	

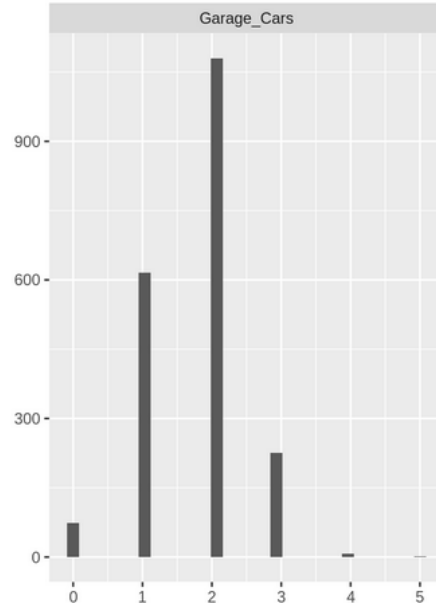
Voici d'autres graphiques non utilisés qui justifient que nous n'ayons pas sélectionné d'autres variables, en raison d'une trop faible (< 50%) corrélation avec le prix de vente :



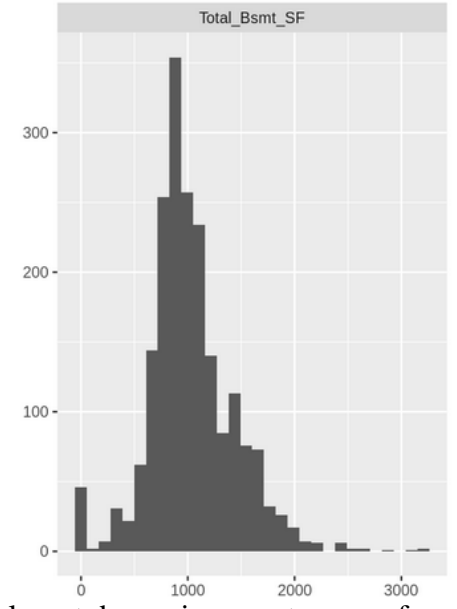
Analysons nos données graphiquement :



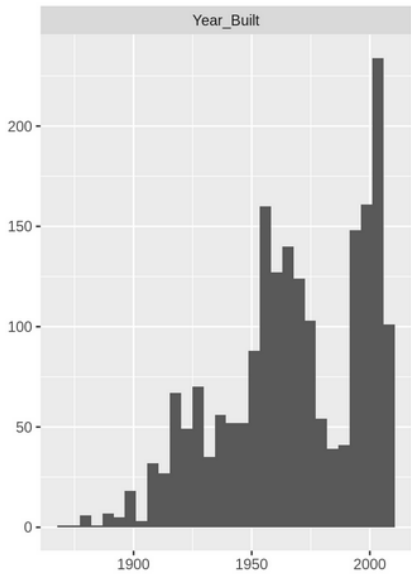
La plupart des maisons ont une (1) à 2 bagnoires (1 étant le plus fréquent). Ces valeurs correspondent bien à l'intuition.



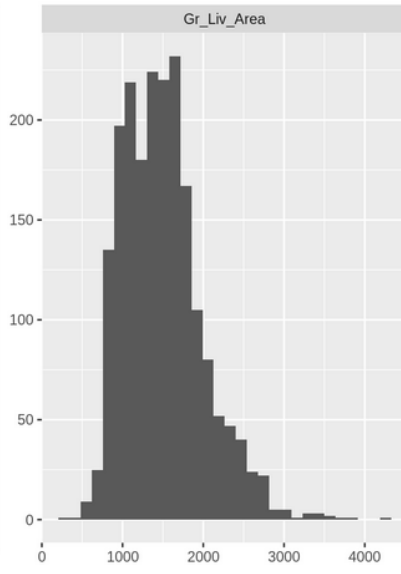
La plupart des maisons ont une (1) à 2 places pour leurs voitures dans leur garage (2 étant le plus fréquent). Cela semble être un schéma assez courant dans les familles.



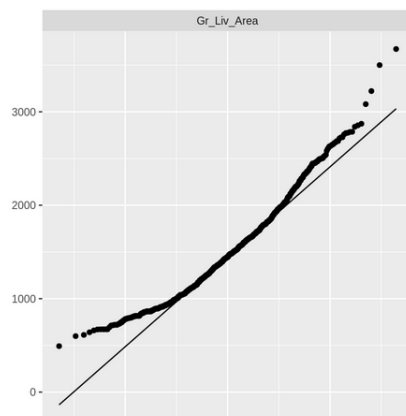
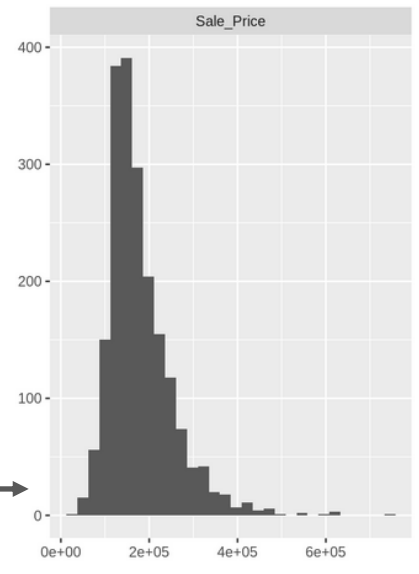
La plupart des maisons ont une surface de sous-sol d'environ 1 000 pieds carrés, avec beaucoup de maisons sans sous-sol. En effet, lorsque l'on a un sous-sol, il est rare qu'il soit extrêmement petit (dans ce cas, on n'en a pas.) ou extrêmement grand.



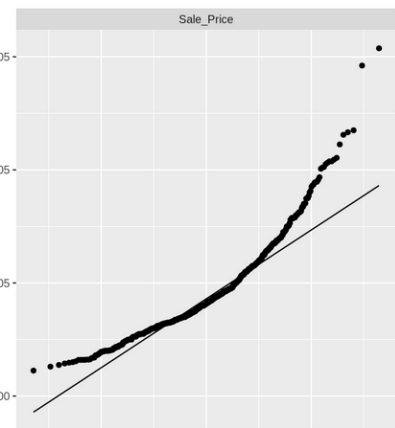
On remarque que malgré un grand nombre de maisons récentes, la plupart datent d'avant 1975. On remarque surtout 2 moments : le nombre de maisons construites après 1945 augmente drastiquement. Cela pourrait s'expliquer par la fin de la 2nd Guerre Mondiale ; le nombre de maisons construites après 1970 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.



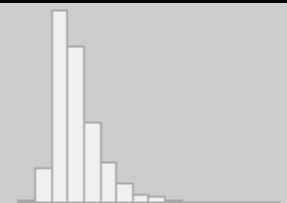
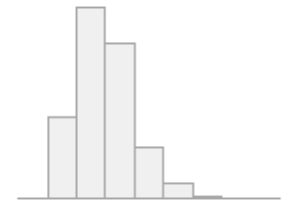

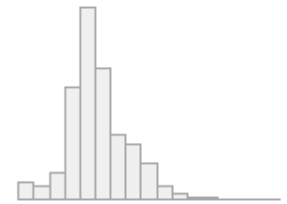
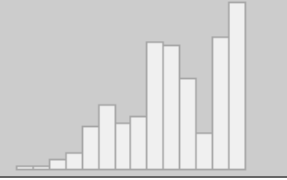

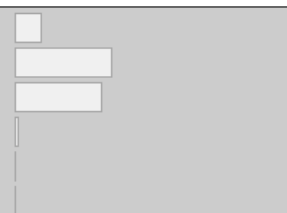

La surface habitable et le prix de vente ont l'air d'avoir une distribution similaire. Une analyse Quantile-Quantile devrait avoir la même allure.



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 estimations.



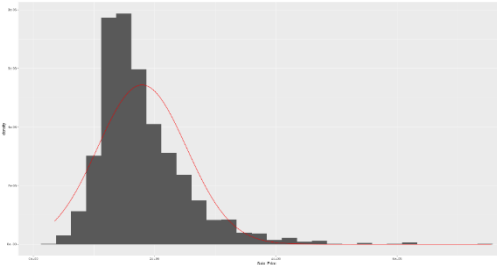
Analysons nos données quantitativement :

Variable	Statistiques	Fréquences	Distribution	Correctes	Manquantes
Sale_Price [integer]	Moyenne (σ) : 179184.6 (73341.2) min \leq médiane \leq max: 35000 \leq 161875 \leq 755000 IQR (CV) : 82387.5 (0.4)	722 valeurs distinctes		2002 (100.0%)	0 (0.0%)
Gr_Liv_Area [integer]	Moyenne (σ) : 1494.3 (495.9) min \leq médiane \leq max: 334 \leq 1445 \leq 4316 IQR (CV) : 650.5 (0.3)	1085 valeurs distinctes		2002 (100.0%)	0 (0.0%)
Garage_Cars [numeric]	Moyenne (σ) : 1.7 (0.7) min \leq médiane \leq max: 0 \leq 2 \leq 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]	Moyenne (σ) : 1031.3 (401.4) min \leq médiane \leq max: 0 \leq 974 \leq 3206 IQR (CV) : 426.8 (0.4)	860 valeurs distinctes		2002 (100.0%)	0 (0.0%)
Year_Built [integer]	Moyenne (σ) : 1967.5 (29.8) min \leq médiane \leq max: 1872 \leq 1968 \leq 2010 IQR (CV) : 46 (0)	113 valeurs distinctes		2002 (100.0%)	0 (0.0%)
Full_Bath [integer]	Moyenne (σ) : 1.5 (0.5) min \leq médiane \leq max: 0 \leq 1 \leq 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 autres]	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%)

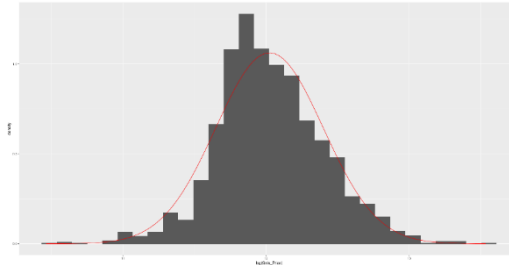
Ainsi, une maison tirée au hasard possèdera très probablement les caractéristiques suivantes :
un prix de 180 000 \$, une surface habitable au-dessus du rez-de-chaussée de 140 m^2 , 2 voitures,
un sous-sol de 90 m^2 , 2 baignoires et sera construite en parpaings à North Ames en 1968.

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 estimations. 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 ($\alpha = 5\%$) que le



$\log(\text{Sale_Price})$ suit une loi normale (p-value > 0.05, donc on ne rejette pas l'hypothèse nulle).

Nous aurons en tout 4 modèles : les 3 premiers sont spécifiés manuellement et le dernier est un compromis entre le R^2 ajusté, le C_p de Mallows et le critère d'information Bayésien (BIC), uniquement pour les variables quantitatives et pour chacun des modèles possibles, cette fois.

1^{er} modèle : $\log(\text{SalePrice}) = \beta_0 + \beta_1 \text{GrLivArea} + u$

Ce modèle n'utilise que GrLivArea pour expliquer le prix, afin de 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 (ce qui est intuitif puisque le prix d'une maison est surtout fonction de sa surface.).

2^{ème} modèle : $\log(\text{SalePrice}) = \beta_0 + \beta_1 \text{GrLivArea} + \beta_2 \text{GarageCars} + \beta_3 \text{TotalBasementSF} + \beta_4 \text{YearBuilt} + \beta_5 \text{FullBath} + \beta_6 \text{Foundation} + \beta_7 \text{Neighborhood} + u$

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

3^{ème} modèle : $\log(\text{SalePrice}) = \beta_0 + \beta_1 \text{MSSubClass} + \dots + \beta_{71} \text{Latitude} + u$

Ce modèle utilise toutes les variables pour expliquer le prix de vente. Son R^2 ajusté permet d'être comparé avec les autres modèles pour vérifier si finalement toutes les variables ne seraient pas importantes. Les coefficients ne seront pas précisés en raison du nombre trop conséquent de variables, mais seulement le R^2 ajusté.

4^{ème} modèle : $\log(\text{SalePrice}) = \beta_0 + \beta_1 \text{LotFrontage} + \beta_2 \text{LotArea} + \beta_3 \text{YearBuilt} + \beta_4 \text{YearRemodAdd} + \beta_5 \text{MasVnrArea} + \beta_6 \text{BsmtFinSF1} + \beta_7 \text{BsmtFinSF2} + \beta_8 \text{BsmtUnfSF} + \beta_9 \text{TotalBsmtSF} + \beta_{10} \text{FirstFlrSF} + \beta_{11} \text{SecondFlrSF} + \beta_{12} \text{BsmtFullBath} + \beta_{13} \text{BedroomAbvGr} + \beta_{14} \text{KitchenAbvGr} + \beta_{15} \text{TotRmsAbvGrd} +$

$$\beta_{16}Fireplaces + \beta_{17}GarageCars + \beta_{18}GarageArea + \beta_{19}WoodDecSF + \beta_{20}ScreenPorch + \beta_{21}PoolArea + u$$

Ces variables ont été choisies en suivant les étapes suivantes :

1. Trouver le meilleur sous-ensemble de variables explicatives (quantitatives) avec l'algorithme Branch & Bound de G. Furnival et W. Wilson [4] via le package « Leaps ».
2. Pour chacun de ces modèles, obtenir son R^2 ajusté, son C_p de Mallows et son BIC.
3. Trouver le modèle qui fait le compromis entre un R^2 ajusté élevé, et un C_p et BIC faible.
4. Utiliser les variables de ce modèle pour les estimations par MCO.

Voici, ci-dessous, les résultats avec en **rouge** le meilleur score par critère, en **bleu** une étendue des meilleurs scores selon chaque critère et en **vert** le choix du modèle qui est le meilleur compromis entre ces 3 critères :

N°	R^2 ajusté	C_p	BIC
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

Il est important de noter que nous travaillons ici uniquement avec une base de données constituée des variables quantitatives retranchées de la base de données initiale. Aussi, bien que nous recherchions 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'avons vu dans l'analyse de corrélation précédente. Nous testerons plus tard cette hypothèse avec l'indicateur VIF. À noter également que nous sélectionnons le meilleur modèle manuellement, en faisant un compromis, mais il existe une méthode plus rigoureuse (par validation croisée) basée sur de l'échantillonnage. Ici, nous choisissons le modèle 21, dont les variables sont explicitées dans le modèle ci-dessus.

4. Résultats

4.1. Statistiques descriptives

Tableau 1 : Prix de vente selon le type de fondation						
	BrkTil	CBlock	PConc	Slab	Stone	Wood
Minimum	35000	52000	59000	39300	65000	143000
Médiane	124950	144100	209000	209000	127950	183000
Moyenne	131251	154592	223457	114267	161235	189750
Maximum	475000	410000	755000	284700	266500	250000

On remarque que, en moyenne, les maisons les moins chères sont en dalle, et les plus chères sont en béton.

Tableau 2 : Répartition des types de fondation par quartier						
	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

On remarque que les maisons les plus vendues à North Ames sont en parpaings.

Tableau 3 : Prix de vente par quartier (quelques quartiers)				
	Minimum	Médiane	Moyenne	Maximum
North_Ames	68000	142000	146904	345000
College_Creek	110000	200500	199779	332000
Old_Town	45000	122000	128156	475000
Edwards	35000	125000	132956	415000
Somerset	176000	245000	248518	468000
Northridge_Heights	214000	326000	345268	615000
Gilbert	115000	184050	189210	377500
Sawyer	62383	135000	137326	219000
Northwest_Ames	127000	185000	194384	306000
Sawyer_West	67500	184900	190508	320000
Mitchell	81500	156225	165515	300000
Brookside	39300	127750	126740	223500
Crawford	90350	196500	199021	381000
Iowa_DOT_and_Rail_Road	40000	118700	113263	212300

On remarque que pour certains quartiers la différence médiane-moyenne n'est pas très grande (par exemple College Creek), mais que pour d'autres quartiers, elle est beaucoup plus importante (par exemple Northridge Heights). En effet, ce dernier est un quartier assez cher, on s'attend donc à ce que le prix d'un petit nombre de maisons dépasse largement la moyenne.

Tableau 4 : Résultats des estimations MCO sur le (log du) prix de vente d'un bien immobilier à Ames				
Variable expliquée : log(SalePrice)	Modèle 1	Modèle 2	Modèle 3	Modèle 4
Variables explicatives :				
GrLivArea	0.0005*** (0.00001)	0.0003*** (0.00001)		
GarageCars		0.07*** (0.007)		
TotalBsmtSF		0.0001*** (0.00001)		
YearBuilt		0.002*** (0.0003)		0.002*** (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***

Garage_Cars				(0.005) 0.04***
Garage_Area				(0.008) 0.0001***
Wood_Deck_SF				(0.00003) 0.00008***
Screen_Porch				(0.00002) 0.0002***
Pool_Area				(0.00005) -0.0001 (0.0001)
Constante	11.14 (0.018)	5.49 (185,8)	-73.14 (41.73)	1.18 (0.3)
N	2002	2002	2002	2002
R ² 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

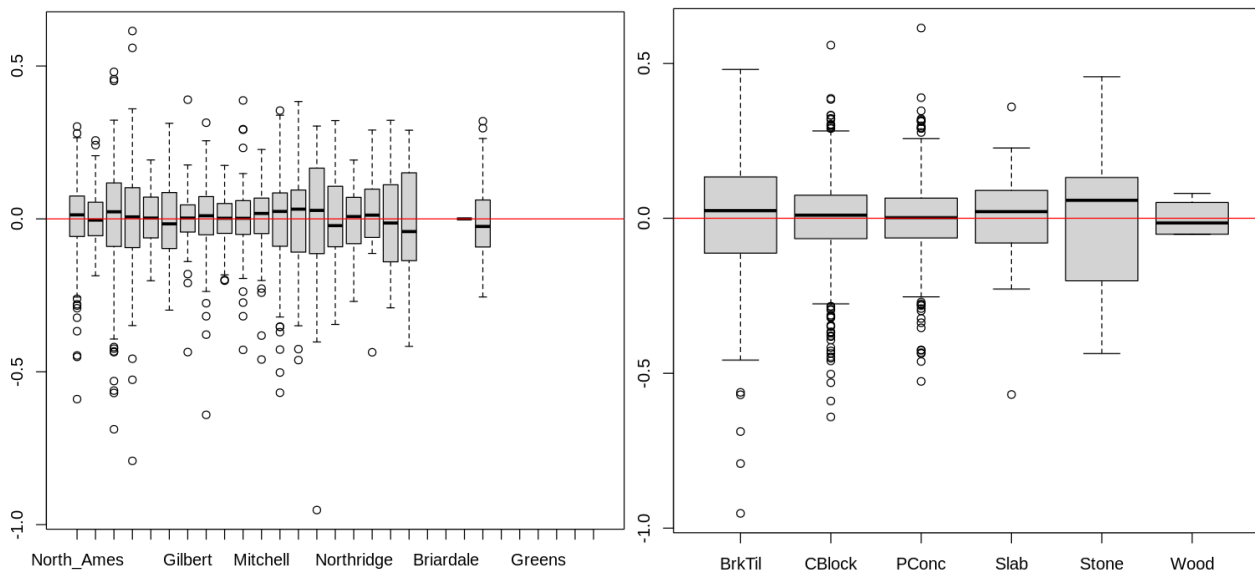
Les erreurs type sont entre parenthèses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

On remarque d'abord des R^2 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](#). 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](#). Une interprétation du modèle 1 corrigé serait de dire que lorsque le nombre de pieds 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 $e^{11.14}$ soit environ 70 000 \$ (cela n'a cependant aucun sens puisque, ici, la surface habitable au-dessus du rez-de-chaussée est toujours supérieure à 0). 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 [5], 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 : peu de personnes veulent d'une maison avec un nombre trop élevé de chambres.

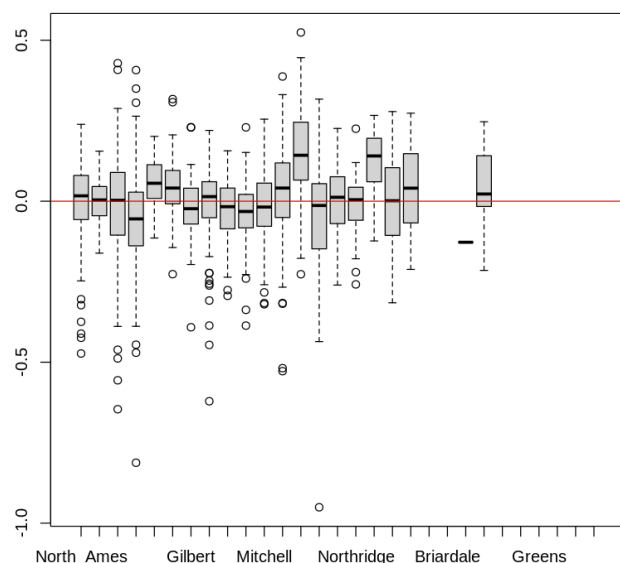
Comme dit précédemment, on remarque un phénomène d'hétéroscédasticité. Nous supposons que les meilleurs candidats pour être les responsables de cette hétéroscédasticité sont les variables « quartier » et « type de fondation », mais ce ne sont que des intuitions pour l'instant.

Voyons de plus près la distribution des résidus en fonction de ces variables pour le modèle 3 :



On remarque en effet que pour le quartier, la variance des résidus n'est pas constante (elle a l'air de baisser vers le quartier Gilbert et augmenter vers le quartier Northridge). Cela peut s'expliquer par le fait que les maisons sont potentiellement construites par lot au sein des quartiers, que les plans et les caractéristiques se ressemblent, mais uniquement au sein du même quartier ou du même lot. Ainsi, un quartier aura plus de chances de trouver des maisons avec les mêmes caractéristiques (et donc avec le même prix) au sein de ce même quartier plutôt qu'au sein d'un autre quartier. Il en va de même pour le type de fondation, puisque le béton étant un

matériau très standard, la variance des résidus sera plutôt faible, alors qu'elle sera plus élevée pour des matériaux moins utilisés comme la pierre. Cette disparité est encore plus vraie pour la variable « quartier » dans le modèle 4 (graphique ci-contre). On y remarque une variance des résidus très peu stable (certains quartiers sont absents en raison des retranchements faits auparavant.).



5. Conclusion

Pour résumer, nous avons analysé les différents facteurs qui rentrent en jeu en ce qui concerne la détermination du prix de vente d'une maison. Nous avons conclu que cette base de données n'est pas tout à fait faite pour des régressions linéaires, a priori, bien que nous le supposions auparavant. En effet, on remarque que peu d'hypothèses sont vérifiées, et donc que beaucoup d'hypothèses importantes ne sont pas vérifiées, comme l'homoscédasticité, la distribution normale des résidus ou encore l'autocorrélation, malgré des spécifications différentes de modèles. 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 éloigné d'un autre quartier. Malgré cela, le R^2 ajusté reste très élevé globalement dans nos modèles. Pour spécifier un modèle meilleur, il faudrait partir de celui avec toutes les variables, enlever chaque variable une par une, puis vérifier s'il y a des différences statistiquement significatives. En effet, avec « seulement » 20 variables, nous arrivons à un R^2 ajusté élevé, 0.88 (contre 0.94 mais 71 variables pour le modèle 3) tout en ayant également un C_p de Mallows 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 le 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, bien que les coefficients ne soient pas biaisés en présence d'hétéroscédasticité, les estimations par MCO ne sont plus les meilleurs (« BLUE »), en raison d'écarts-types (et donc de statistiques de Student) peu fiables. À 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](#)]
- [2] Bureau of Labor Statistics, U.S. Department of Labor. Consumer Expenditures – 2020 [[lien](#)]
- [3] LADWEIN, Richard et EREM, IAE. Le jugement de typicalité comme heuristique de choix: approche comparative. In: RA Perterson A. Jolibert et A. Strazzieri, editeurs, Proceeding of the International Research Seminar. 1995. p. 351. [[lien](#)]
- [4] Furnival, G. M., & Wilson, R. W. (2000). Regressions by leaps and bounds. Technometrics, 42(1), 69-79. [[lien](#)]
- [5] Current Population Survey, 2020 ASEC Technical Documentation [[lien](#)]

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