Elora GODARD

Chloé LEYRIS

SAE 4.EMS.01 : Prédiction et explication du prix d'un bien à partir d'un jeu de donnée

IUT2 Université Grenoble Alpes

Département STID

Table des matières

Introduction

Ce rapport vise à élaborer un modèle de données concernant des propriétés résidentielles. Le jeu de données étudié comprend un ensemble de variables détaillant diverses caractéristiques des maisons, notamment le prix, la surface habitable, le nombre de chambres, salles de bains, garages, étages, ainsi que la présence de climatisation, de chauffage, etc.

L'objectif principal de cette étude est de comprendre et de modéliser les variations des prix des maisons en fonction d'une combinaison de ces paramètres. Autrement dit, nous cherchons à identifier les facteurs qui influent le plus sur la valorisation d'une propriété résidentielle à travers une modélisation linéaire multiple.

Dans cette étude nous allons réaliser une série d'analyses statistiques et de modélisation des données, visant à identifier les relations entre les différentes variables et à établir un modèle pour estimer les prix des maisons. Pour cela, nous réaliserons une analyse exploratoire qui identifiera les tendances et corrélations. Puis, nous procéderons à la construction du modèle. Enfin, nous présenterons une conclusion à notre analyse.

Description des données

Nous cherchons à identifier les facteurs qui influent le plus sur la valorisation d'une propriété résidentielle. Parmi les 12 variables explicatives disponibles dans le jeu de données, nous distinguons :

- une variable quantitative continue, « area ». Il s'agit de la surface du bien en pieds². Cette variable présente beaucoup de modalités dites rares (1 individu par modalités). Nous avons donc décidé de regrouper les surfaces en 3 classes : entre 0 et 4000 pieds², entre 4000 et 7000 pieds² et plus de 7000 pieds². Ainsi, on obtient une nouvelle variable qualitative ordinale, « area_classe ».

- 4 variables quantitatives discrètes : « bedrooms » qui correspond au nombre de chambres, « bathrooms » qui correspond au nombre de salles de bain, « stories » qui correspond au d'étages et « parking » qui correspond au nombre de places de parking.

- 6 variables binaires, avec des modalités "yes" et "no". Parmi ces variables, il y a « mainroad » indiquant si la maison est à proximité ou non d'une route principale, « guestroom » indiquant la présence ou non d'une chambre d'amis, « basement » indiquant la présence ou non d'un sous-sol, « hotwaterheating » indiquant la présence ou non d'un système de chauffage, « airconditioning » indiquant la présence ou non d'une climatisation et enfin « prefarea » indiquant si la localisation du bien est bonne ou non.

- une variable qualitative ordinale « furnishingstatus » à 3 modalités « furnishing », « semi-furnishing » et « unfurnishing ». Cette variable indique le niveau d'ameublement du bien.

Nous choisissons la variable « price » comme variable à expliquer dans notre modèle. sous l'hypothèse que les différentes variables du jeu de données contribuent à la variation du prix du bien.

Dans notre analyse, nous garderons les noms de base du jeu de données, c'est-à-dire les noms anglais.

Analyse Univariée

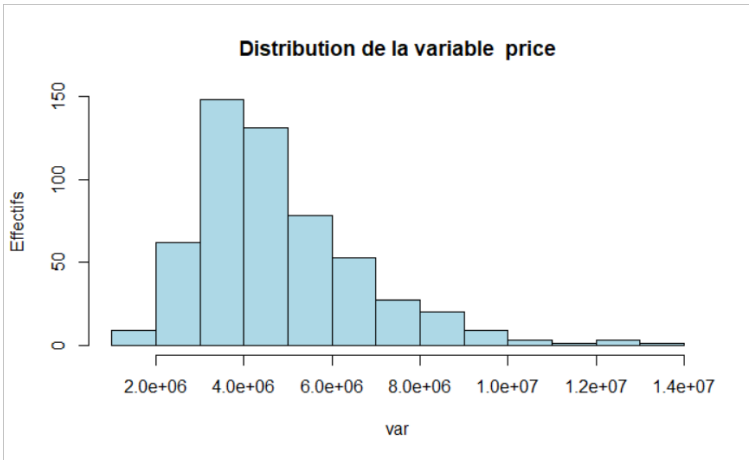
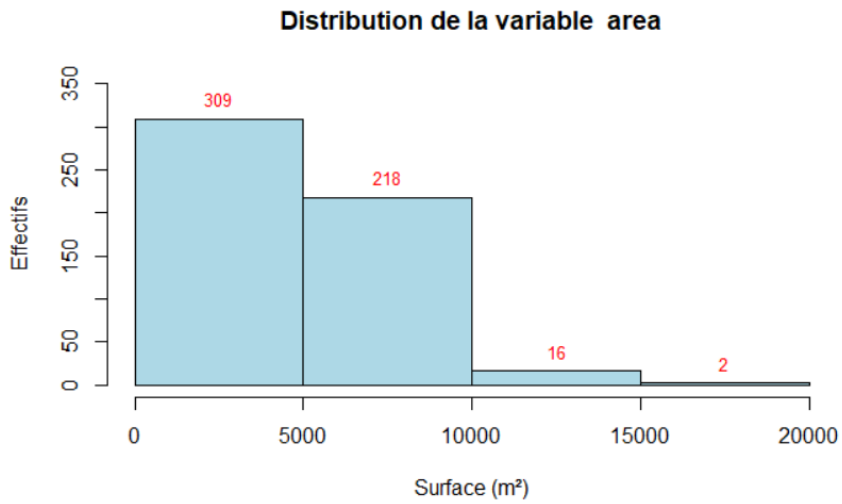


Figure 1: histogramme de la variable Price

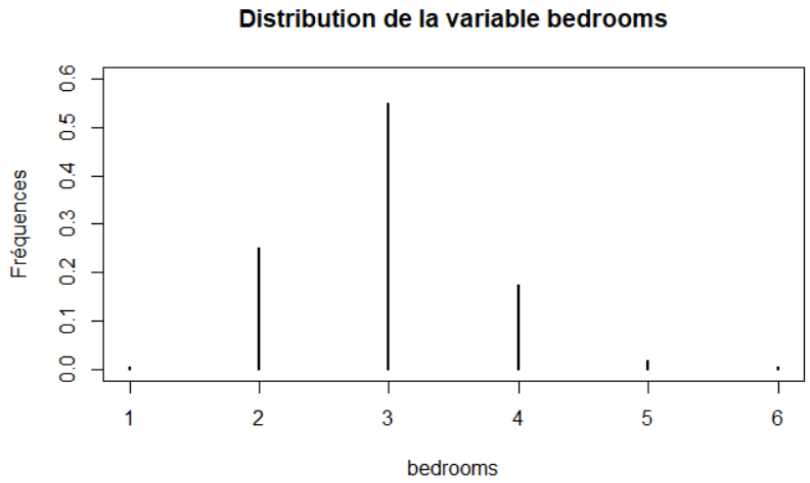
Dans la figure ci-dessus, nous présentons la distribution de la variable « price », que nous utiliserons comme variable explicative Y dans notre analyse. Nous constatons que le prix des maisons varie, allant de 1 750 000 $ pour la propriété la moins chère à 13 300 000 $ pour la plus chère. La moyenne du prix des maisons dans notre échantillon est de 4 766 729 $. De plus on observe une très grande variance dans notre échantillon.

Il est également intéressant de noter que la distribution des prix des maisons semble suivre une loi normale. Cela suggère que la majorité des maisons dans notre échantillon se situent autour de la moyenne.

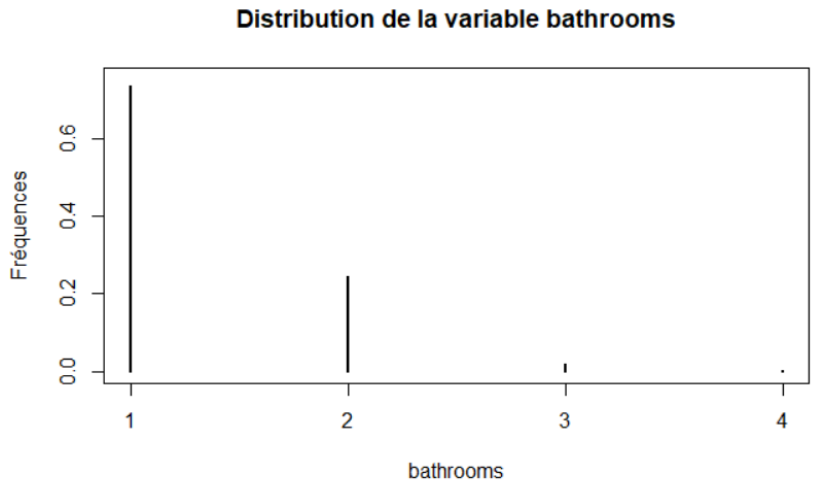


La variable « surface » correspond à la superficie des maisons. Pour mieux représenter les différentes modalités, nous avons décidé de les regrouper en classes. En effet, la répartition des surfaces est très inégale, la majorité des maisons ayant une superficie comprise entre 1650

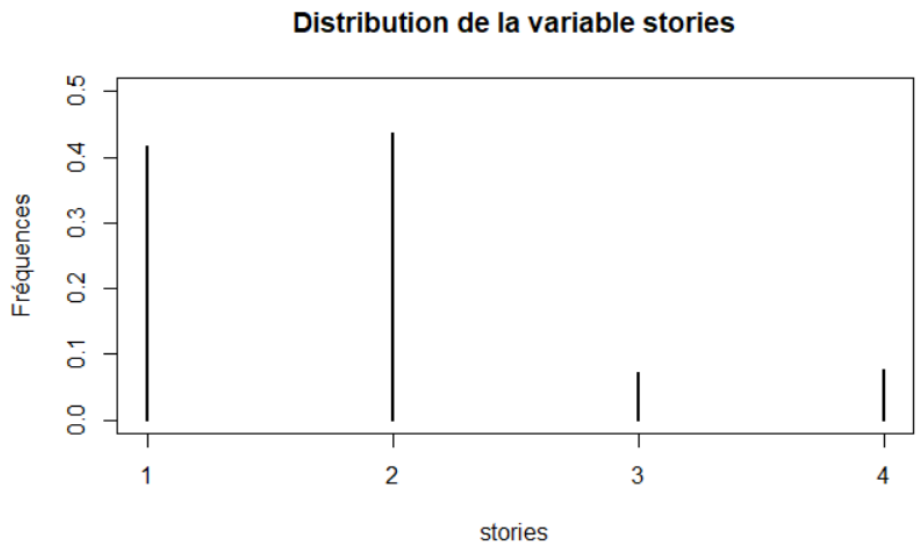
et 5000 m². À l'autre extrémité, nous trouvons seulement deux maisons ayant une superficie de plus de 15000 m², la plus grande étant de 16200 m². Ainsi, la moyenne de la superficie est de 5151 m² et la variance est de 4 709 512 m²²



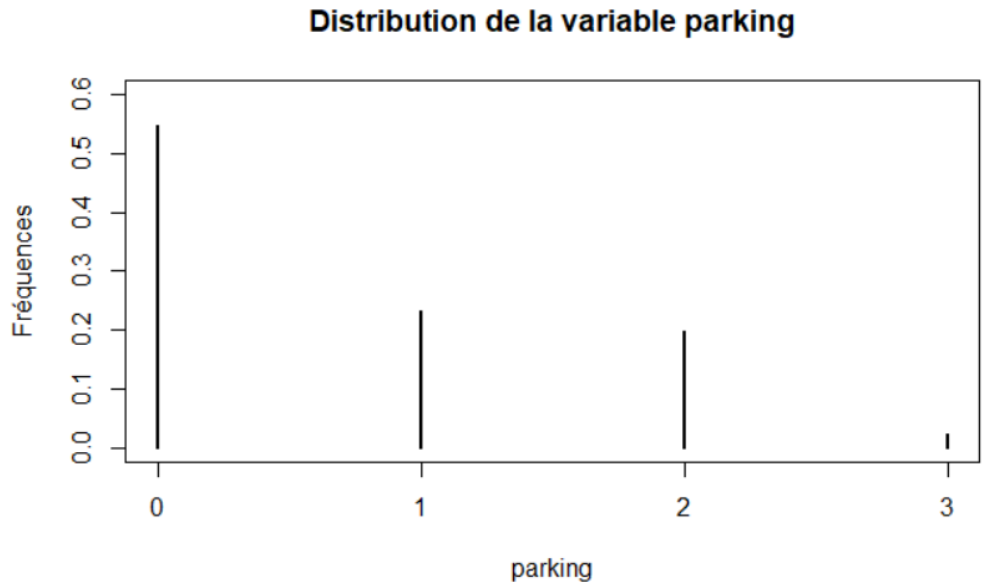
Sur le diagramme à barres, nous observons la répartition de la variable « bedrooms », représentant le nombre de chambres dans une maison donnée. L'étendue de cette variable est de 5, allant de 1 à 6 chambres. La modalité la plus courante est celle de 3 chambres, représentant plus de 55 % des maisons dans notre échantillon. En revanche, les modalités comptant le moins de maisons sont celles de 1 et 6 chambres.



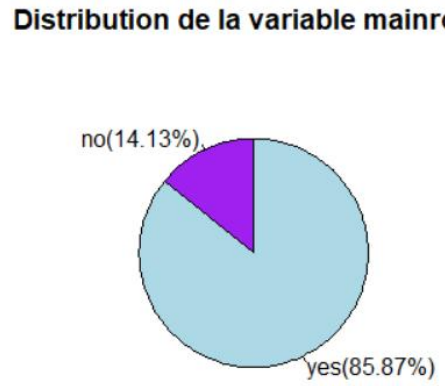
Ce graphique présente la répartition de la variable "salles de bain", qui représente le nombre de salles de bain dans une maison. Nous observons que 401 maisons de notre échantillon possèdent une seule salle de bain. En examinant les données, il apparaît que plus le nombre de salles de bain augmente, moins il y a de maisons concernées par cette modalité.



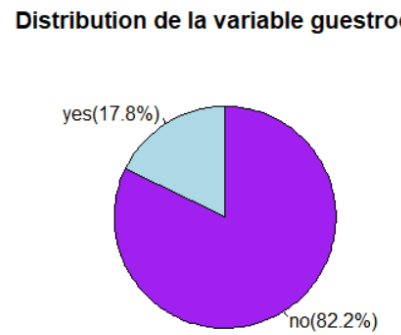
Dans cette représentation, nous étudions la variable "stories", qui indique le nombre d'étages dans une maison. Cependant, il convient de préciser que, par exemple, "2 stories" correspond en français à un rez-de-chaussée plus un étage. Ainsi, nous constatons que 45% de l'échantillon possède un seul étage. De plus, plus de 40% des maisons de l'échantillon ne possèdent aucun étage. Les deux autres modalités sont plus rares car peut de maison ont plus d'un étage.



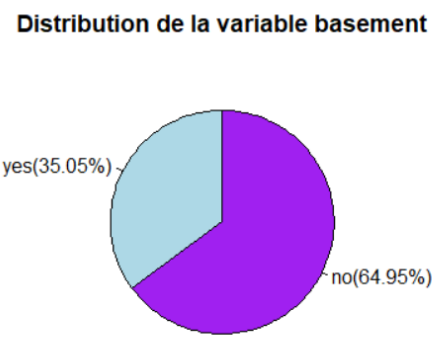
Nous avons représenté la variable "parking" sous forme d'un diagramme à barres. Cette variable correspond au nombre de places de parking associées à chaque bien immobilier. Il est remarquable que plus de la moitié de la population d'échantillon ne dispose d'aucune place de parking. De plus, nous observons une tendance selon laquelle plus le nombre de places de parking est élevé, moins cette modalité est courante. Par exemple, une centaine de maisons ont une ou deux places de parking associées, tandis que seulement 12 maisons en possèdent trois.



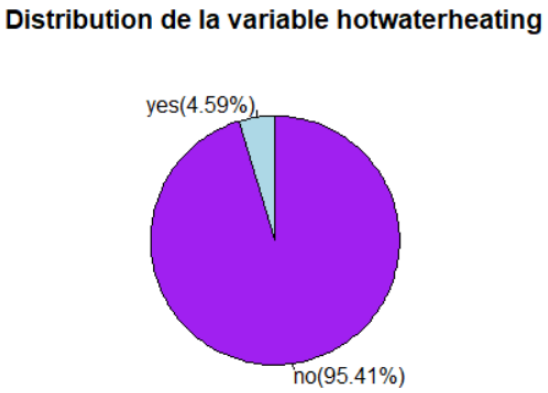
Nous avons dressé le diagramme circulaire pour la variable binaire "mainroad", qui indique si le bien est situé sur une route principale ou non. Nous constatons que plus de 85% de la population d'échantillon est effectivement située sur une route principale.



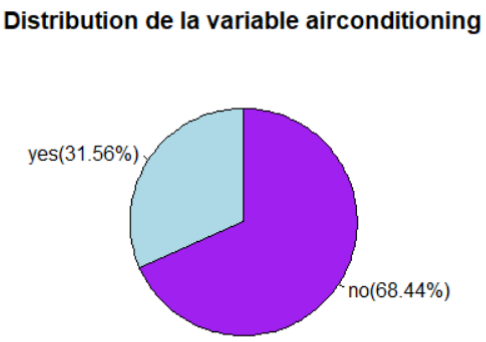
La variable "guestroom" est représentée ci-dessus, elle indique si le bien dispose d'une chambre d'amis. Nous constatons que plus de 80 % des maisons ne possèdent pas cette chambre supplémentaire.



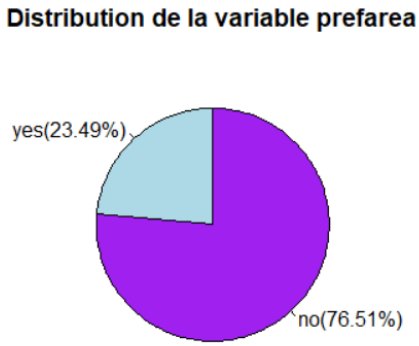
Nous avons représenté la variable « basement », qui indique si la maison est accompagnée d'un sous-sol. Nous constatons que 35 % des biens possèdent un sous-sol.



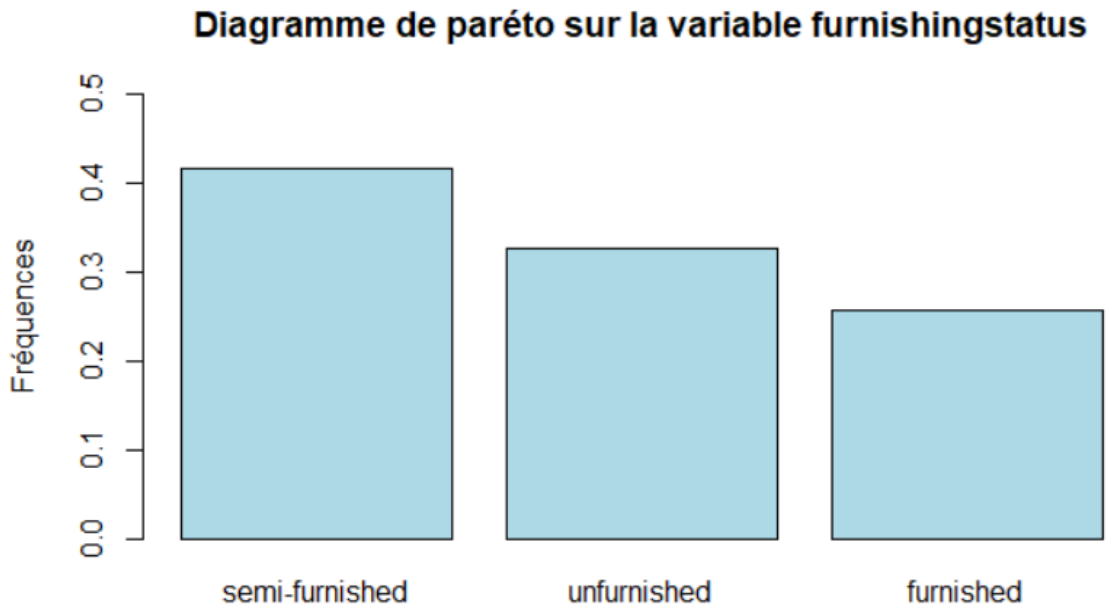
Nous réalisons un diagramme circulaire pour représenter la variable « chauffage à eau chaude », qui correspond aux systèmes de chauffage où de l'eau chauffée circule à travers un réseau de tuyaux vers des radiateurs ou des unités de traitement d'air, qui libèrent ensuite la chaleur dans l'air ambiant. Cependant, il n'est pas obligatoire d'en installer dans une maison. Plus de 95 % des biens n'en possèdent pas.



Nous avons représenté la variable « airconditioning » sous forme de diagramme circulaire. Cette variable nous renseigne sur la présence ou non de la climatisation dans le bien. Près de 70 % des maisons ne possèdent pas de climatisation.



La variable « prefarea » indique si la maison est située dans un quartier prisé ou non. Nous l'avons représentée ci-dessus sous forme d'un diagramme circulaire. Seulement 1 maison sur 4 se trouve dans un quartier désirable.



Enfin, nous présentons la variable qualitative ordinale « furnishingstatus ». Cette variable comporte trois niveaux de meublage dans la maison, allant de non meublée à entièrement meublée. La modalité mi-meublé est la plus répandue, tandis que les maisons entièrement meublées sont plus rares.

Etude des corrélations

Nous souhaitons étudier les corrélations entre les variables explicatives.

bedrooms	bathrooms	stories	mainroad	guestroom	basement	hotwaterhea	aircondition	parking	prefarea	furnishingsta	area_classe
1	0,3739	0,4086	-0,012	0,0805	0,0973	0,046	0,1606	0,1393	0,079	-0,1232	0,1577
0,3739	1	0,3262	0,0424	0,1265	0,1021	0,0672	0,1869	0,1775	0,0635	-0,1436	0,1884
0,4086	0,3262	1	0,1217	0,0435	-0,1724	0,0188	0,2936	0,0455	0,0444	-0,1047	0,1107
-0,012	0,0424	0,1217	1	0,0923	0,044	-0,0118	0,1054	0,2044	0,1999	-0,1567	0,2874
0,0805	0,1265	0,0435	0,0923	1	0,3721	-0,0103	0,1382	0,0375	0,1609	-0,1183	0,1487
0,0973	0,1021	-0,1724	0,044	0,3721	1	0,0044	0,0473	0,0515	0,2281	-0,1128	0,0649
0,046	0,0672	0,0188	-0,0118	-0,0103	0,0044	1	-0,13	0,0679	-0,0594	-0,0316	-0,0023
0,1606	0,1869	0,2936	0,1054	0,1382	0,0473	-0,13	1	0,1592	0,1174	-0,1505	0,2604
0,1393	0,1775	0,0455	0,2044	0,0375	0,0515	0,0679	0,1592	1	0,0916	-0,1775	0,3681
0,079	0,0635	0,0444	0,1999	0,1609	0,2281	-0,0594	0,1174	0,0916	1	-0,1077	0,2021
-0,1232	-0,1436	-0,1047	-0,1567	-0,1183	-0,1128	-0,0316	-0,1505	-0,1775	-0,1077	1	-0,1644
0,1577	0,1884	0,1107	0,2874	0,1487	0,0649	-0,0023	0,2604	0,3681	0,2021	-0,1644	1

On peut considérer qu'il y a de l'interaction entre 2 variables lorsque la valeur absolue de la corrélation est supérieure à 0.5. On peut voir en jaune dans le tableau XXXX la plus haute variable en valeur absolue. On conclut donc qu'aucune variable ne répond à ce critère. Ainsi, il n'y a pas de corrélations des variables entre elles.

Ensuite, on souhaite étudier la corrélation des variables explicatives avec notre variable à expliquer, Prix.

bedrooms	bathrooms	stories	mainroad	guestroom	basement	hotwaterheating	airconditioning	parking	prefarea	furnishingstatus	area_classe
0,3665	0,5175	0,4207	0,2969	0,2555	0,1871	0,0931	0,453	0,3844	0,3298	-0,3047	0,5318

On constate alors dans le tableau XXX que 2 variables ont une corrélation supérieure à 0.5. Il s'agit des variables « bathrooms » et « area ». On peut donc dire que ces variables sont plutôt corrélées à la variable. Globalement, on remarque que les corrélations sont assez hautes.

L'étude de ces corrélations nous donnent une première idée des variables qui présentent un effet sur la variable prix. Nous allons ensuite vérifier cela dans la suite de notre analyse.

Stratégie de modélisation :

Dans un premier temps nous allons faire l'étude sur le modèle complet. Ainsi, nous pourrons voir les variables qui ont un effet sur la variable à expliquer. Le modèle complet s'écrit sous la forme :

Y = B0 + B1*bedrooms + B2*bathrooms + B3*stories + B4*mainroad + B5*guestroom + B6*basement + B7*hotwaterheating + B8*airconditioning + B9*parking + B10*prefarea + B11*furnishingstatus + B12*area_classe

Par la suite, nous réalisons un codage sur les variables qualitatives. Nous commençons par convertir toutes nos variables en types « factor » afin de faciliter l'analyse. Ces codages nous permettrons d'obtenir des matrices de design pour chacune de nos variables qualitatives. On réalise donc le codage par défaut de R pour les variables binaires. Ce codage

indique que le coefficient B0 correspond à la moyenne de la modalité 1 (ici la modalité « no »). Ensuite, le coefficient B1 correspond à la différence entre la moyenne de la modalité 2 (modalité « yes ») et celle de la modalité 1. Si nous avions une troisième modalité, nous aurions soustrait sa moyenne à celle de la modalité 1. Ainsi, les matrices obtenues par ce codage n'auront qu'une seule colonne car les variables ont 2 modalités.

Pour les variables qualitatives ordinales, le codage est légèrement différent. En effet, ce dernier compare la moyenne entre 2 groupes successifs. Ainsi, B0 correspond toujours à la moyenne de la première modalité. B1 représente donc la différence entre la moyenne de la modalité 2 et la modalité 1. A la différence du codage précédent, B2 représente la différence entre la moyenne de la modalité 3 et la modalité 2. Les matrices de design qui découlent de ces codages auront donc 2 colonnes car les variables ont 3 modalités.

Ainsi, lorsqu'on regroupe toutes les matrices, on obtient une matrice de design que nous utiliserons par la suite dans notre modèle. (tableau XXX)

Yes_mainroad	Yes_airconditioning	Yes_basement	Yes_guestroom	Yes_hotwaterheating	Yes_prefarea	furnished	unfurnished	moins_4000	plus_7000
1	1	0	0	0	1	1	0	0	1
1	1	0	0	0	0	1	0	0	1
1	0	1	0	0	1	0	0	0	1
1	1	1	0	0	1	1	0	0	1
1	1	1	1	0	0	1	0	0	1
1	1	1	0	0	1	0	0	0	1
1	1	0	0	0	1	0	0	0	1
1	0	0	0	0	0	0	1	0	1
1	1	1	1	0	1	1	0	0	1
1	1	0	1	0	1	0	1	0	0
1	1	1	0	0	1	1	0	0	1
1	0	1	1	1	0	0	0	0	0
1	1	0	0	0	1	0	0	0	0
1	0	0	0	1	0	1	0	1	0
1	0	0	0	0	1	0	0	0	1
1	0	1	0	0	0	0	0	0	0
1	1	1	1	0	1	0	1	0	0
1	1	0	0	0	0	1	0	0	1
1	1	0	1	0	0	1	0	0	0

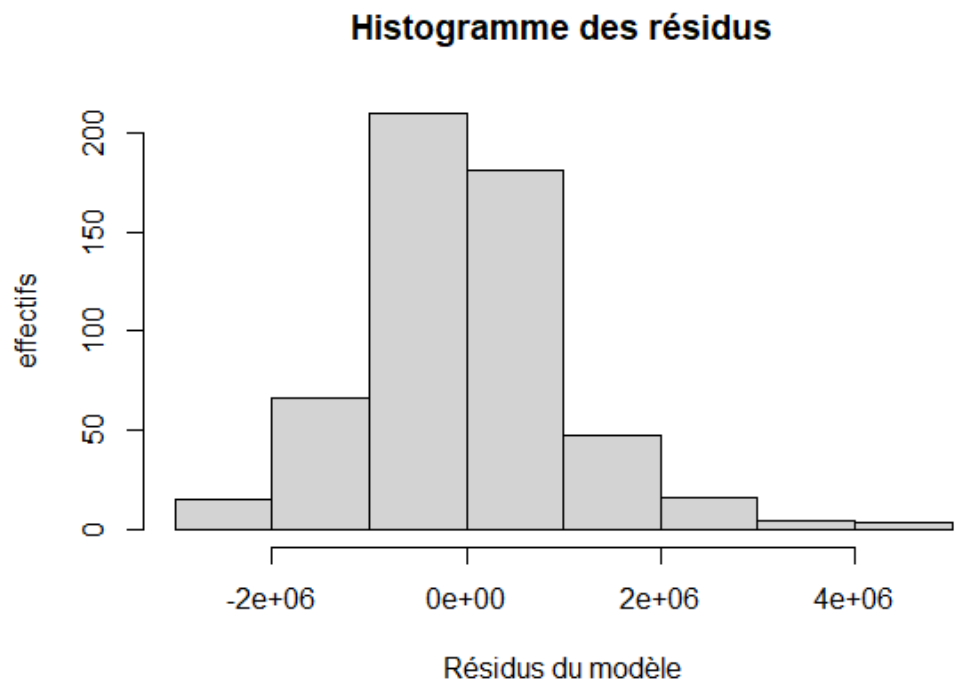
(20 première lignes de la matrice de design)

Par la suite, l'ANOVA permet de déterminer l'effet de nos variables dans le modèle. Pour cela, nous observons les p-values du modèle dont la valeur est inférieure à 0.05 (tableau XXX).

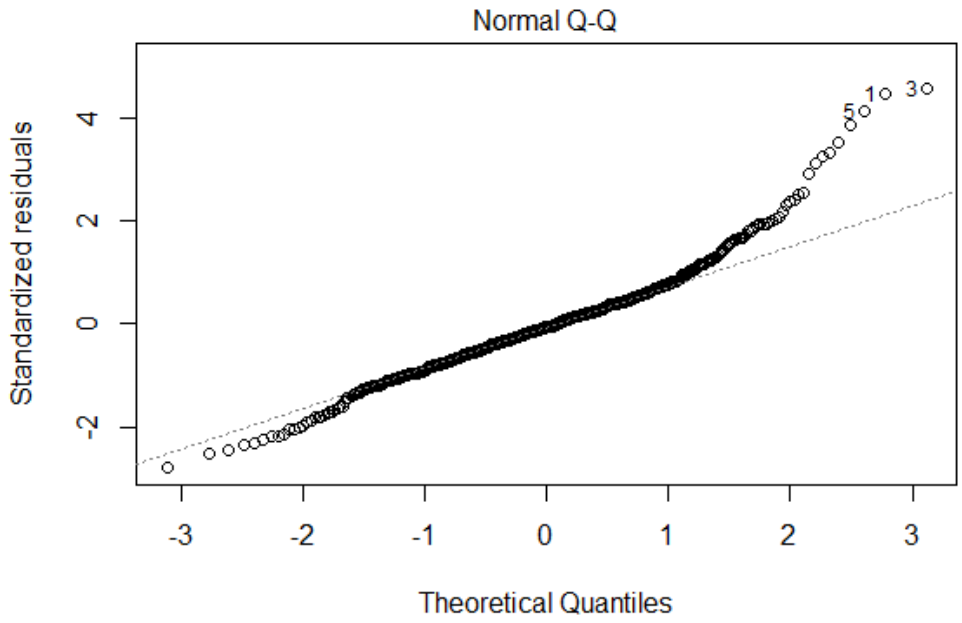
	pvalue
bedrooms	0,02959413
bathrooms	4,18E-16
stories	4,93E-11
parking	6,36E-05
mainroad_yes	2,97E-05
airconditioning_yes	8,70E-14
basement_yes	0,00015943
guestroom_yes	0,01011077
hotwaterheating_yes	0,00033415
prefarea_yes	9,78E-10
furnished	0,74629745
unfurnished	1,21E-08
area_moins4000	5,80E-14
area_plus7000	2,75E-05

On constate dans le tableau XXX, le codage des variables qualitatives auxquelles nous avons ajouté les modalités qui ne correspondent pas à la modalité de référence dans le nom des variables. Cela explique par exemple les « yes » à la fin de certains noms de variables. De plus, on constate que presque toutes les p-values sont inférieures à 0.05, sauf la variable « furnished ». Cette variable correspond à une modalité de la variable « furnishingstatuts » que nous avons codée précédemment. Or, ici, on estime qu'il est important de garder cette variable car l'autre modalité « unfurnished » a un effet conséquent sur le modèle. Ainsi, toutes les variables du modèle ont un effet certain sur le prix.

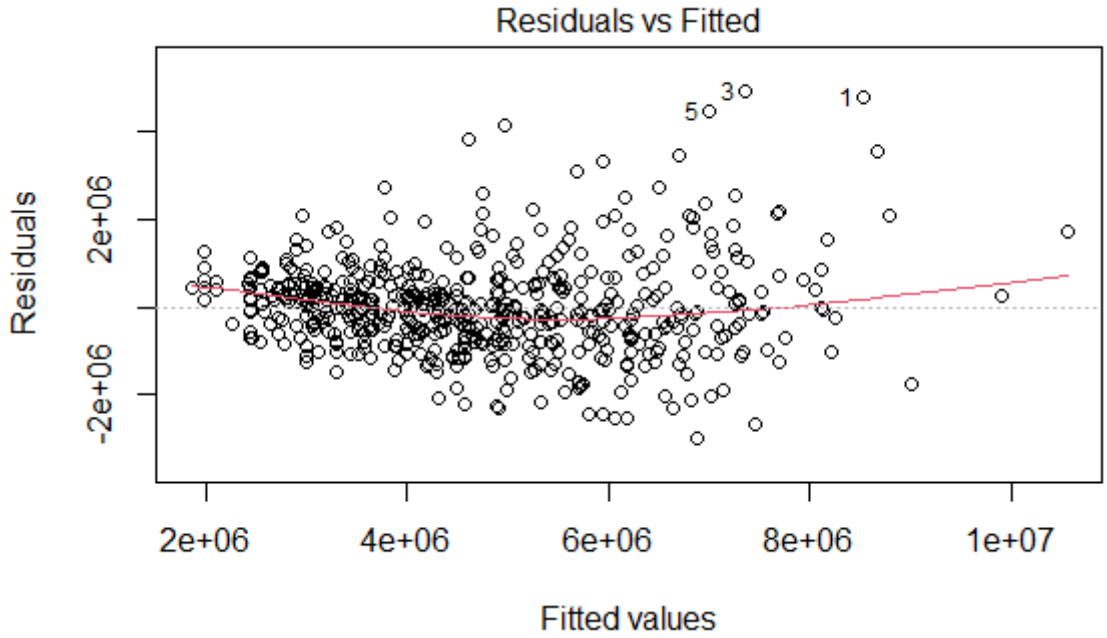
En revanche, nous ne pouvons pas encore valider le modèle. En effet, par la suite il est important de réaliser une étude des résidus afin d'améliorer les données si nécessaire (en effectuant une transformation). Pour cela nous allons tout d'abord observer l'histogramme des résidus (figure XXX).



On constate alors que la répartition des résidus semble plus ou moins suivre une loi normale, mais ce n'est pas une évidence. Afin de confirmer nos doutes, nous observons ensuite le QQplot du modèle (figure XXX).



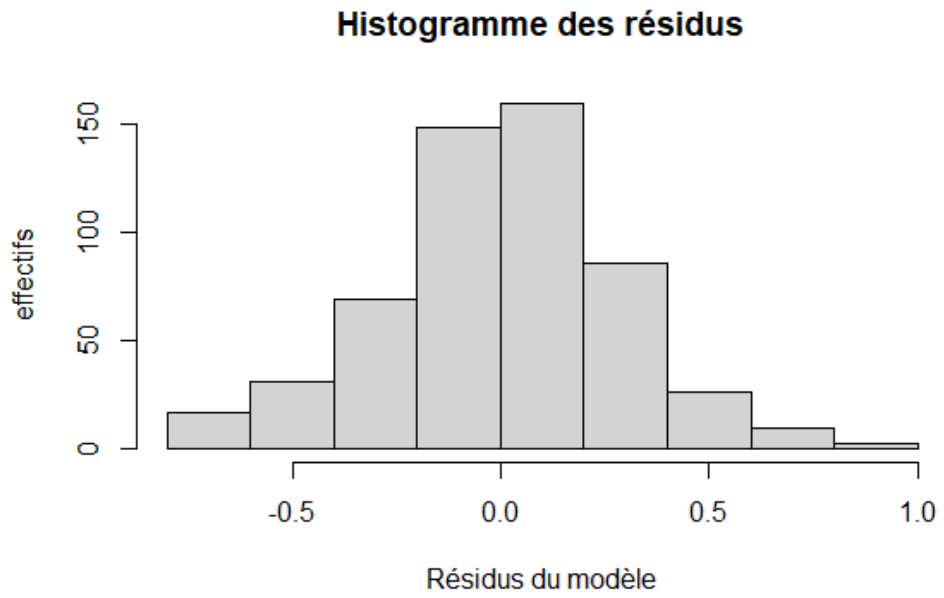
On remarque alors que cette dernière s'éloigne fortement de la droite d'Henry à droite du graphique. Cela nous confirme que le modèle ne semble pas suivre une loi gaussienne. De plus, si on observe la figure XXX, on constate une forme d'entonnoir dans le nuage de point, montrant aussi que le modèle n'est pas gaussien.



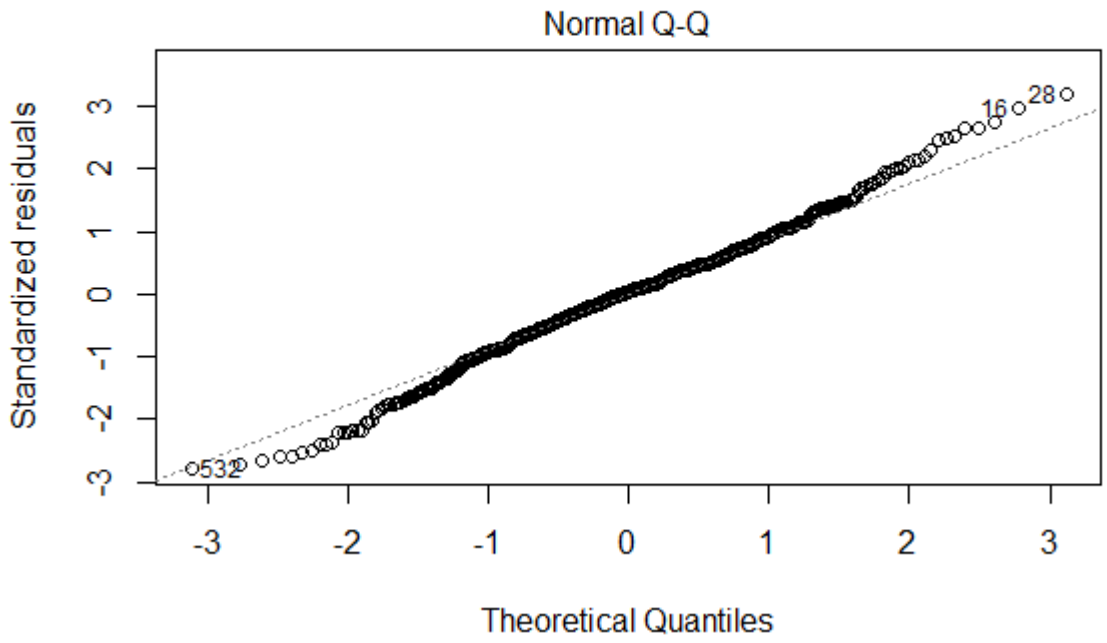
Transformation de la variable « prix »

Pour contrer cela, nous allons effectuer une transformation sur la variable y. Nous allons utiliser la fonction « Boxcox ». Il s'agit d'une généralisation de la transformation logarithmique. Le principe de cette fonction est d'appliquer une puissance λ à y. λ permet de

maximiser la log-vraisemblance. On obtient alors un λ = 0.020 environ (différent de 0) pour le modèle. On peut ainsi transformer y avec la formule : (yλ − 1)/λ. Ainsi, nous continuerons l'analyse avec ce y transformé.



On observe de nouveau l'histogramme des résidus (figure XXX) et on remarque qu'il semble que la transformation a fonctionnée. En effet, l'histogramme semble suivre une loi normale. On confirme alors avec le QQPlot et la droite d'Henry (figure XXX).



On peut alors affirmer la normalité des résidus avec la figure XXX, ce qui est nécessaire pour que la validation du modèle. Par la suite, on étudie l'hypothèse d'homoscédasticité sur le modèle. Pour cela, il est conseillé d'utiliser un test de Bartlett. Or, avec les modifications que nous avons faites sur nos données, il n'est pas possible de le faire. Ainsi, on utilise une fonction « check_heteroscédasticité » du package « performance ». Cette fonction nous donne alors une p-value d'environ 0,85. Ainsi, elle est largement supérieure à 0.05, on peut donc valider l'hypothèse d'homoscédasticité.

On peut ensuite observer de nouveau les effets de nos variables sur le modèle avec le y transformé (tableau XXX).

	pvalue
bedrooms	0,03001737
bathrooms	3,01E-16
stories	4,54E-11
parking	5,97E-05
mainroad_yes	3,15E-05
airconditioning_yes	7,84E-14
basement_yes	0,00016969
guestroom_yes	0,00998125
hotwaterheating_yes	0,00032103
prefarea_yes	9,19E-10
furnished	0,76166413
unfurnished	1,56E-08
area_moins4000	6,92E-14
area_plus7000	2,38E-05

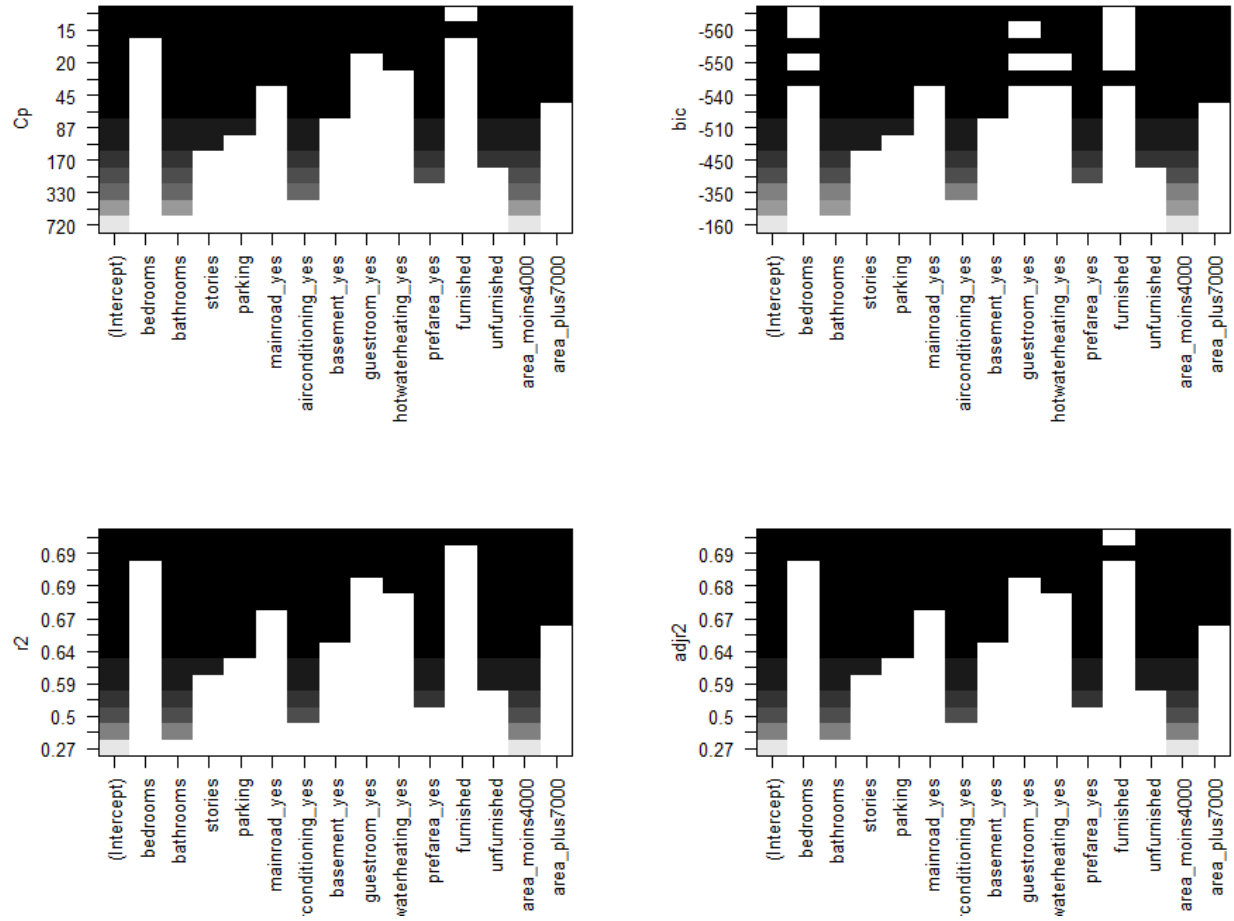
On constate que la conclusion est la même que pour le y non transformé. En effet, on constate que presque toutes les p-values sont inférieures à 0.05, sauf la variable « furnished ». Comme pour le modèle précédent, on garde cette variable.

Ensuite, nous évaluons la multicolinéarité du modèle en calculant le VIF(*variance inflation factor*) pour chacune des variables quantitatives du modèle. Cette dernière permettra de nous dire certaines variables sont expliquées par d'autres, dû à la multicolinéarité entre elles.

	VIF
bedrooms	1,31206611
bathrooms	1,23849091
stories	1,25513124
parking	1,04097749

On constate alors dans le tableau XXX que tous les VIF sont inférieurs à 5. Cela signifie qu'il n'y a pas de multicolinéarité évidente. Ainsi, les variables ne sont pas expliquées entre elles. Cela confirme l'hypothèse dans l'analyse des corrélations. En effet, nous avions déjà constaté très peu de corrélations entre les variables explicatives.

Pour finir sur la vérification des variables qui ont un effet sur notre prix, nous utilisons des plots qui dépendent de 4 paramètres : le Cp, le BIC, le R2 et le R2 ajusté. On obtient alors la figure XXX.



Le meilleur modèle correspond à la ligne la plus haute des graphiques. SI le carré est noir, cela signifie que la variable est significative. Ainsi, on arrive à la même conclusion que précédemment. En effet, la modalité « furnished » de la variable furnishingstatus est la seule qui n'est pas prise sur la plupart des graphiques. Or, comme dit précédemment, nous allons la garder. On voit aussi sur un des graphiques que la variables « bedrooms » n'est pas significative pour le paramètre du BIC. On décide tout de même de la garder car dans les analyses précédentes elle semblait significative.

Ainsi, on conclut que notre modèle complet, avec toutes les variables de départ, est le meilleur modèle que nous pouvions obtenir, sans interactions entre les variables.

Interprétation du modèle parcimonieux

Par la suite, il est important d'analyser les différents coefficients de nos variables.

variable	nom_coeffs	coefficient	intervalle_inf	intervalle_sup	dans_intervalle
(Intercept)	B0	17,05907922	16,92188096	17,19627747	TRUE
bedrooms	B1	0,042022571	0,004080519	0,079964624	TRUE
bathrooms	B2	0,231771092	0,17783525	0,285706935	TRUE
stories	B3	0,114982262	0,081396762	0,148567762	TRUE
parking	B4	0,064450469	0,033161548	0,09573939	TRUE
mainroad_yes	B5	0,159619882	0,084935304	0,23430446	TRUE
airconditioning_yes	B6	0,224250285	0,1668753	0,28162527	TRUE
basement_yes	B7	0,110967836	0,05341039	0,168525282	TRUE
guestroom_yes	B8	0,091223456	0,021920173	0,16052674	TRUE
hotwaterheating_yes	B9	0,215191284	0,098463095	0,331919472	TRUE
prefarea_yes	B10	0,190884741	0,130750438	0,251019045	TRUE
furnished	B11	-0,009449013	-0,070618765	0,051720739	TRUE
unfurnished	B12	-0,167193982	-0,224377964	-0,110009999	TRUE
area_moins4000	B13	-0,226850249	-0,284756962	-0,168943536	TRUE
area_plus7000	B14	0,15863916	0,08554273	0,23173559	TRUE

On constate dans le tableau XXX que lorsqu'aucun facteur n'est pris en compte dans le prix d'une maison, ce dernier est de 17.05. Ensuite, on constate que le nombre de chambre, de place de parking, la présence d'une chambre d'amis élèvent légèrement le prix. Le nombre de salle de bain, d'étages, la présence d'une route principale, de climatisation, d'un sous-sol, de chauffage, d'une bonne localisation et d'une surface supérieure à 700, font augmenter le prix plus fortement. En revanche, une surface inférieure à 4000, une maison ayant des meubles fournis ou non, font baisser le prix des maisons plus ou moins fortement.

On observe par la suite les intervalles de confiance. On constate que tous nos coefficients sont compris dans les intervalles (TRUE dans la dernière colonne) avec un risque d'erreur de 5%. On peut donc conclure que tous nos coefficients sont cohérents avec les intervalles.

Prédiction

Enfin, nous allons utiliser le modèle dans un cadre de prévisions de données. Etant donné qu'il y a beaucoup d'individus dans le jeu de données de départ (545), nous décidons de découper notre échantillon pour la prévision.

Dans un premier vecteur on tire aléatoirement 70% du jeu de données et on met les 30% restants dans un autre vecteur. Par la suite, on applique le modèle complet pour le jeu de données à 70%. Ensuite, on applique ce nouveau modèle au jeu de données de 30% que l'on veut prédire (tableau XXX).

y_test	y_predict.fit	y_predict.lwr	y_predict.upr	différence.fit
12250000	10929974,11	8741274,113	13118674,11	1320025,886
12250000	7445681,662	5326251,193	9565112,131	4804318,338
12215000	8819716,273	6695266,595	10944165,95	3395283,727
11410000	6969360,27	4839877,541	9098842,999	4440639,73
9800000	7795493,3	5658895,847	9932090,752	2004506,7
8750000	5428234,553	3274509,885	7581959,22	3321765,447
8400000	7359779,887	5178931,137	9540628,637	1040220,113
8400000	6920978,446	4778982,601	9062974,292	1479021,554
7980000	8318801,143	6195386,74	10442215,55	-338801,1428
7700000	6698732,687	4578800,181	8818665,193	1001267,313
7560000	7295593,135	5186064,052	9405122,218	264406,8647
7420000	7765827,298	5629225,327	9902429,27	-345827,2981
7420000	5472385,342	3353005,175	7591765,508	1947614,658
7350000	7670715,694	5547360,522	9794070,865	-320715,6936
7343000	6223536,436	4101156,499	8345916,372	1119463,564
7245000	9197576,035	7056812,272	11338339,8	-1952576,035
7210000	8407635,588	6262318,459	10552952,72	-1197635,588
7210000	7448527,854	5320276,665	9576779,043	-238527,8536
7070000	5324649,672	3208639,025	7440660,319	1745350,328
6720000	5976037,675	3850280,975	8101794,376	743962,3247

La colonne y_test correspond à notre échantillon, c'est-à-dire les valeurs que nous essayons de prédire. Y_predict.fit présente les valeurs prédites grâce au modèle élaboré. Avec la commande predict, nous créons un intervalle de confiance. Enfin, la dernière colonne correspond à epsilon, qui est l'erreur générée par le modèle lorsqu'il prédit une valeur. Nous remarquons que toutes nos valeurs prédites appartiennent à l'intervalle de confiance, ce qui signifie qu'elles sont proches de la réalité. Nous notons également que l'ordre de grandeur de nos erreurs est en centaines de milliers, voire en millions. Nous expliquons ces valeurs élevées par le fait que nos prix de maison sont exprimés en plusieurs millions de dollars.

En résumé, notre modèle de prédiction montre une bonne performance avec toutes les valeurs prédites se situant dans l'intervalle de confiance, bien que les erreurs associées soient souvent élevées, ce qui peut être expliqué par la nature des prix des maisons dans notre ensemble de données.

Conclusion

En conclusion, ce rapport s'est concentré sur l'élaboration d'un modèle de données pour les propriétés résidentielles, en explorant un jeu de données comprenant diverses caractéristiques telles que le prix, la surface, le nombre de chambres, de salles de bain, etc. L'objectif principal était de comprendre les variations des prix des maisons en fonction de ces paramètres à travers une modélisation linéaire multiple. Nous avons réalisé une analyse exploratoire des données, identifiant les tendances et corrélations, puis construit un modèle complet. Malgré des erreurs parfois élevées, nos prédictions sont cohérentes avec la réalité, témoignant de la robustesse du modèle dans l'estimation des prix des maisons. Notre modèle complet étant dès le début satisfaisant nous n’avons pas eu à effectuer le modèle d’ANCOVA.

Pour aller plus loin, une piste d'ouverture intéressante serait d'explorer l'incorporation de données supplémentaires dans notre modèle, telles que les données économiques locales, les taux d'intérêt hypothécaires, ou même des données environnementales comme la proximité des transports en commun ou des espaces verts. Cette approche pourrait enrichir notre compréhension des facteurs influençant les prix des propriétés résidentielles et permettre le développement de modèles encore plus précis et informatifs pour les décideurs immobiliers, les acheteurs potentiels et les investisseurs.