Projet Econometrie

Lancelot Ravier, Othmane Laaraki, Onur Ozdemir

December 31, 2023

1 Introduction

Le marché de l'immobilier est un élément central de l'économie et de la vie quotidienne de nombreuses personnes. Ce marché influence non seulement la stabilité financière des ménages, mais il joue également un rôle essentiel dans le développement économique régional.
Dans ce contexte, l'étude du prix des biens immobiliers revêt une importance cruciale pour comprendre les dynamiques économiques et sociales d'une région donnée. Ce projet empirique se consacre à l'analyse du marché immobilier du comté de King, situé dans l'État de Washington, et a pour objectif l'analyse des déterminants principaux du prix des biens immobiliers dans cette région.
Le comté de King, qui englobe les villes de Seattle et de Bellevue, est l'une des régions métropolitaines les plus dynamiques et diversifiées des États-Unis. Cette région est caractérisée par une économie en croissance constante, une industrie technologique prospère, une scène culturelle riche et une qualité de vie prisée. En conséquent, le marché immobilier de King County est particulièrement actif.
Cette étude vise donc à répondre à la question suivante :
Quels-sont les facteurs les plus significatifs influençant les prix des biens immobiliers dans le comté de King et comment peuvent-ils être quantifiés en terme d'impact sur le prix de vente des biens ?
Les objectifs de cette étude sont les suivants : premièrement, nous allons analyser une base de données ainsi que les tendances entre les variables, puis nous allons constituer et tester différents modèles par la méthode des Moindres Carrés Ordinaires (MCO). Pour finir, certains tests seront menés sur notre modèle le plus performant afin de discuter de la significativité et de la pertinence des résultats obtenus, puis de la robustesse de ce modèle dans sa globalité.

2 Présentation de la database

La base de données analysée ci-dessous provient de l'administration du comté de King. Elle contient 21613 observations de biens immobiliers vendus dans le comté de King, entre Mai 2014 et Mai 2015, ainsi que la date, la référence de la vente et 18 caractéristiques descriptives de ces biens vendus.

```
RangeIndex: 21613 entries, 0 to 21612
Data columns (total 21 columns):

 #   Column         Dtype
---  ------         -----
 0   id             numero d'identification du bien vendu
 1   date           date de vente du bien
 2   price          prix de vente du bien (en $ 2015)
 3   bedrooms       nombre de chambres
 4   bathrooms      nombre de salles de bains / WC (0.5 par WC et 1 par salle de bain)
 5   sqft_living    superficie habitable (en pieds carré)
 6   sqft_lot       superficie totale du bien (en pieds carré)
 7   floors         nombre d'étages
 8   waterfront     =1 si le bien est situé en zone cotière (vue sur la cote), 0 sinon
 9   view           notation de la qualité de la vue, de 0 (min) à 4 (max)
 10  condition      notation de l'état général du logement, de 1 (min) à 5 (max)
 11  grade          notation de la qualité de construction et de design de l'appartement, de 1 (min) à 13 (max)
 12  sqft_above     superficie intérieure (en pieds carré) du bien située au-dessus du rez-de-chaussée
 13  sqft_basement  superficie intérieure (en pieds carré) du bien située en-dessous du rez-de-chaussée
 14  yr_built       année de construction du logement
 15  yr_renovated   année de la dernière rénovation du logement (0 si non rénové)
 16  zipcode        code postal
 17  lat            latitude (du sud au nord)
 18  long           longitude (d'ouest à est)
 19  sqft_living15  la superficie (en pieds carré) habitable des 15 biens voisins les moins éloignés
 20  sqft_lot15     la superficie (en pieds carré) totale des 15 biens voisins les moins éloignés
```

Figure 1: Description des variables

3 Traitement des données

3.1 Outliers

Une première analyse des statistiques descriptives nous montre que certaines observations contiennent des valeurs erronées :

	Min	Max
bedrooms		33,00
bathrooms	0,0000	

Figure 2: Anomalies

Pour les observations dont bathrooms = 0, il n'existe pas de bien construit sans WC. Pour ce qui est de l'observation dont bedrooms = 33, le nombre de chambres ne semble pas cohérent face à la superficie habitable du bien et à son nombre de salles de bain. Ainsi, ces observations, considérées comme erronées, ont été supprimées.

3.2 Traitement des données

Afin de permettre une meilleure analyse de cette base de données, certaines variables ont été créées puis ajoutées à la dataframe.

- La variable age a été créée pour permettre une meilleure interprétation quantitative de l'age des logements.
- La variable renovated (dummy) a été créée, permettant de distinguer les biens rénovés des autres biens
- La variable centre_ville a été créée, permettant de distinguer les biens situés en ville (=1) des autres biens (=0) selon leur code postal

#	Column	Non-Null Count	Dtype
22	age	age du bien lors de la vente (issue de : sales_yr - yr_built)	
23	renovated	=1 si le bien a subi une ou plusieurs rénovations, =0 sinon	
24	centre_ville	=1 si le bien est situé dans les villes de Seattle ou Bellevue, =0 sinon	

Figure 3: Variables créées

3.3 Données manquantes

Après analyse, aucune donnée manquante n'est à constater dans la dataframe.

4 Statistiques descriptives et analyse des relations entre les données

D'après les principaux sites de vente de biens immobiliers aux États-unis, les principaux facteurs influençant le prix des biens sont la localisation géographique du bien, la superficie habitable, l'état du bien, la qualité de la vue et l'age du bien. L'objectif de cette analyse est ici d'analyser les statistiques descriptives des variables sélectionnées, d'identifier les tendances et relations entre les caractéristiques des biens vendus et la variable expliquée : le prix de vente des biens.

4.1 Statistiques descriptives

	Moyenne	Médiane	Minimum	Maximum
price	540130	450000	78000	7770000
l_price	13,048	13,017	11,264	15,857
sqft_living	2080,1	1910,0	370,00	13540
view	0,23424	0,00000	0,00000	4,0000
condition	3,4096	3,0000	1,0000	5,0000
lat	47,560	47,572	47,156	47,778
long	-122,21	-122,23	-122,52	-121,32
age	43,317	40,000	-1,0000	115,00

	Ec. type	C.V.	Asymétrie	Ex. kurtosis
price	3,6712e+005	0,67969	4,0256	34,598
l_price	0,52643	0,040346	0,43052	0,68970
sqft_living	918,16	0,44140	1,4725	5,2491
view	0,76631	3,2715	3,3965	10,896
condition	0,65048	0,19078	1,0359	0,52032
lat	0,13856	0,0029133	-0,48507	-0,67652
long	0,14075	0,0011517	0,88453	1,0508
age	29,377	0,67818	0,46913	-0,65816

Figure 4: Statistiques descriptives pour les variables sélectionnées

D'après la Figure 3 :

price Le prix moyen des biens immobiliers est d'environ 540130$, avec une médiane de 450000$. Les prix des biens varient de 78000$ à 7700000$. L'écart-type est élevé (367120$), indiquant une variabilité significative dans le prix de vente des biens. La distribution est asymétrique à la droite (asymétrie = 4.02656) et présente une queue lourde (excès de Kurtosis = 34.598), ce qui suggère que la plupart des biens se situent dans la tranche inférieure de prix, avec quelques biens au prix significativement plus élevé.

l_price Le logarithme du prix moyen est de 13.048, avec une médiane très proche de la moyenne, indiquant une distribution plus symétrique pour les prix transformés en logarithme. L'écart type (0.52643) est moindre que pour les prix non transformés, ce qui montre un degré de variabilité moindre par rapport aux prix réels. Les valeurs d'asymétrie et d'excès de kurtosis sont plus proches de celles d'une distribution normale, ce qui suggère que la transformation logarithmique des prix normalise leur distribution. Ainsi, on préfèrera l'utilisation de cette variable comme variable expliquée dans nos modèles.

sqft_living La taille moyenne est de 2080.1 pieds carré, avec une médiane légèrement inférieure à 1910 pieds carrés. Les superficies des biens vont de 370 à 13540 pieds carrés. La distribution est également asymétrique à la droite (asymétrie = 1.4725) et leptokurtique (excès de kurtosis = 5.2491).

view La note moyenne de la vue est comprise entre 0 et 1, mais la médiane est ce 0, ce qui indique que la plupart des propriétés ont une vue de qualité basse. La distribution est fortement asymétrique (asymétrie = 3.3965) et présente un excès de kurtosis élevé (10.896), décrivant la qualité de la vue comme une caractéristique rare et donc précieuse.

condition La note moyenne de l'état des biens est comprise entre 3 et 4, avec une médiane de 3. La plupart des propriétés sont donc en état moyen à bon, comme l'indique le faible écart-type (0.65048).

lat / long La latitude et longitude donnent les valeurs de constitution de la cartographie du comté de King et des positions des différents biens vendus. Pour la latitude, l'écart-type est de 0.13856, ce qui montre une faible variabilité dans la latitude des propriétés. La distribution de la latitude est légèrement asymétrique à gauche (asymétrie = -0.48507) et présente un excès de Kurtosis négatif (-0.67652), indiquant une distribution relativement uniforme avec une quantité moindre de valeurs extrêmes que dans une distribution normale. L'écart-type de la longitude est de 0.14075, indiquant également une faible variabilité de la longitude des propriétés. La distribution de la longitude est asymétrique à droite (asymétrie = 0.88453) et présente un léger excès de Kurtosis (1.0508), ce qui suggère une concentration de valeurs vers une extrémité de la distribution.

En résumé, ces statistiques indiquent une grande variété de prix et de caractéristiques comme la superficie et la qualité de la vue suggèrent un marché immobilier diversifié. L'asymétrie et l'excès de kurtosis dans les prix et la superficie habitable indiquent la présence de biens de grande valeur en tant qu'exceptions, ce qui est typique sur les marchés immobiliers. Pour ce qui est de la géolocalisation des biens, ces statistiques indiquent que les biens immobiliers vendus dans le comté de King sont concentrés dans une zone géographique relativement restreinte, avec une variabilité limitée, en termes de latitude et de longitude, cette intuition étant consolidée par la légère asymétrie et les valeurs de Kurtosis pour les deux variables qui suggèrent des concentrations spécifiques de propriétés dans certaines zone. La transformation logarithmique des prix par rapproche la distribution d'une distribution normale, ce qui est bénéfique pour la modélisation économique et justifie de ce fait l'utilisation de cette variable transformée comme variable explicative.

4.2 Matrice de corrélation

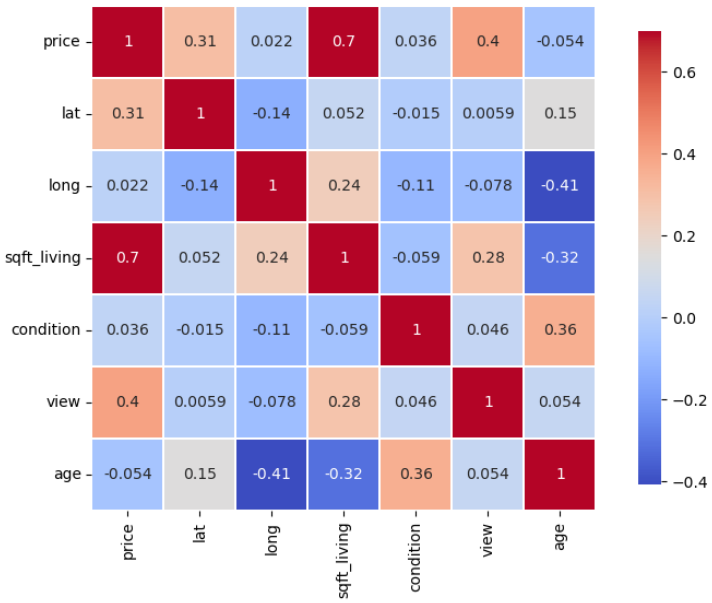
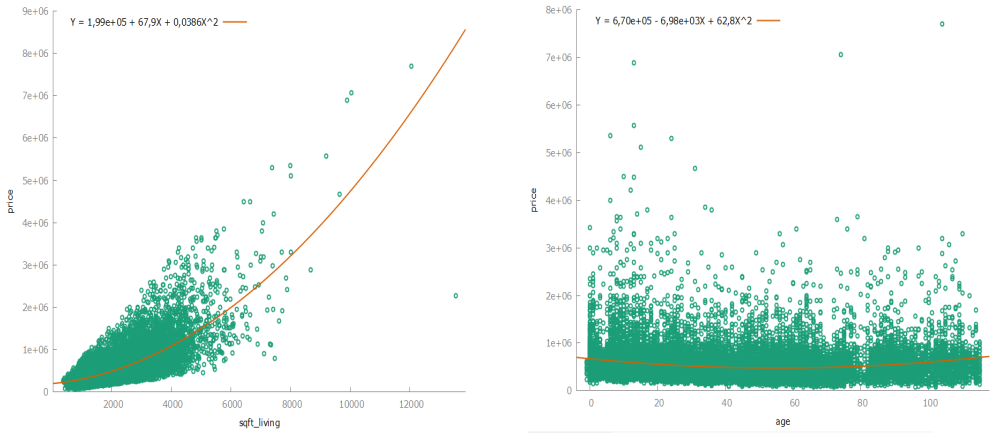


Figure 5: Matrice de corrélation

A partir de cette matrice, nous pouvons conclure qu'il n'existe pas de corrélations critiques (supérieur à .75) entre les variables explicatives, ce qui réduit le risque d'overfitting (sur-ajustement, réduisant les performances des modèles lorsqu'ils sont appliqués à d'autres bases de données ou d'autres situations) de nos futurs modèles.
Aussi, nous pouvons identifier les corrélations entre les variables explicatives et le prix :

- Il existe une corrélation forte entre le prix et la superficie habitable.
- Il existe une corrélation modérée entre le prix et la qualité de la vue / la latitude
- Il n'existe pas de corrélation significative entre le prix et la longitude / l'état du bien / l'age du bien

4.3 Relations entre les variables

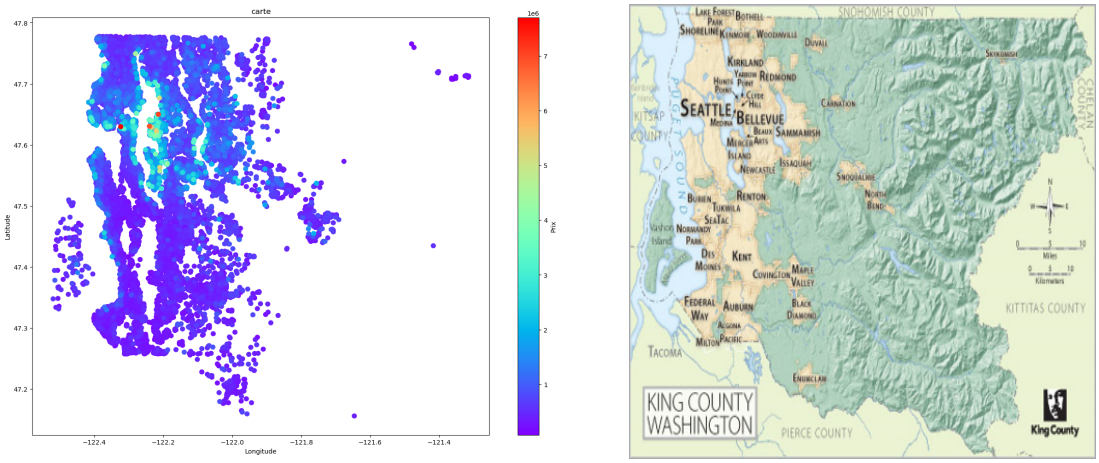


(a) sqft_living vs prix

(b) age vs prix

Figure 6: Graphiques de relation au prix et droites d'ajustement quadratique

A partir de ces deux graphique ci-dessus (et de leur courbe d'ajustement quadratique respectives), nous pouvons observer une relation non linéaire entre le prix et la surface habitable / l'age du bien. Les courbes de tendance semblent indiquer une relation quadratique. L'ajout d'un terme au carré pour la variable sqft_living et age serait alors justifié.



(a) longitude vs latitude (b) Comté de King, Washington

Figure 7: Carographie des biens vendus dans le comté de King

Selon ce graphique, le prix des logements augmente lorsqu'on se déplace du sud au nord, et baisse lorsque l'on se déplace d'ouest à est. Ainsi, ce graphique montre que les biens vendus dans les villes de Seattle et Bellevue sont vendus à un prix supérieur aux biens vendus en zone rurale, et que la pluspart des biens vendus se concentrent sur une zone géographique précise.

5 Modèle à estimer

L'objectif de ce projet est d'estimer le prix de vente des biens immobilier dans le comté de King, à Washington (États-Unis). Ainsi, après avoir analyser les différentes relations entre les variables et, en se basant sur les principaux facteurs influençant le prix de vente des biens fournis par les sites de vente immobilier des États-unis, nous allons tenter d'estimer par les MCO le log du prix en fonction de la localisation géographique , de la superficie habitable, de la qualité du bien , de la qualité de la vue, de l'état du bien et de son age :

$$l\_price = \beta_0 + \beta_1 \cdot lat + \beta_2 \cdot long + \beta_3 \cdot sqft\_living + \beta_4 \cdot condition + \beta_5 \cdot view + \beta_6 \cdot age + \varepsilon \quad (1)$$

L'ensemble des paramètres du modèle sont interprétés comme des semi-élasticités : ceux-ci donnent donc des variations de $(\beta * 100)\%$ du prix lorsque la variable explicative augmente d'une unité.

view / condition Les paramètres liés à view et condition devraient être de signe positif : le prix d'un bien augmente lorsqu'il est en meilleur état et qu'il offre une meilleure vue.

sqft_living Le paramètre lié à sqft_living devrait être positif : les biens plus spacieux ont un prix de vente plus élevés. Cependant, l'ajout d'un pied carré à un studio n'aura pas le même effet sur le prix que l'ajout d'un pied carré dans une villa. Ainsi, il serait pertinent d'ajouter un terme au carré pour capturer cet effet non linéaire.

age Le paramètre lié à age devrait être lui négatif car lorsqu'un logement vieillit, il se dégrade et perds de sa valeur. Cependant, certains biens prennent de la valeur avec le temps pour diverses raisons (rénovations, fluctuations sur les marchés immobilier, biens très anciens) : il serait alors pertinent d'inclure un terme au carré pour capturer cette relation non-linéaire.

lat / long Les paramètres liés à latitude et longitude devraient être de signe positif et négatif. En effet, comme on peut le voir sur la Figure 7, les biens situés au nord ont tendance à être vendus plus cher que les biens situés au sud, ce qui est de même pour les biens situés à l'ouest par rapport aux biens situés à l'est.

6 Résultats d'estimation

6.1 Modèle 1 : MCO

```
Modèle 1: MCO, utilisant les observations 1-21602
Variable dépendante: l_price

               coefficient  p. critique
  ---------------------------------------
  const        -71,2690       1,46e-268  ***
  lat            1,55287      0,0000     ***
  long          -0,0774632    1,21e-06   ***
  sqft_living    0,000370797  0,0000     ***
  condition      0,0528027    2,36e-056  ***
  view           0,105866     1,64e-305  ***
  age            0,000443920  1,28e-07   ***
```

Figure 8: Modèle 1

D'après les résultats d'estimation, tout les paramètres sont significatifs au seuil de 1%. Cependant, ces résultats de test sont valable sous hypothèse de variance de l'erreur constante. En présence d'heteroscedasticité, la variance de l'erreur n'est plus constante pour toutes les observations, ce qui aura une influence sur la matrice des variances-covariances, et les tests qui en découlent (ex : t de student). Nous avons donc réalisé un test de White pour détecter une éventuelle heteroscedasticité dans notre modèle de régression :

```
Test de White pour l'hétéroscédasticité
MCO, utilisant les observations 1-21602
Variable dépendante: uhat^2

  R2 non-ajusté = 0,151046

Statistique de test: TR^2 = 3262,898158,
avec p. critique = P(Khi-deux(27) > 3262,898158) = 0,000000
```

Figure 9: Test de White (Modèle 1)

Le test de White montre la présence d'heteroscedasticité au seuil de 1%. Afin de rectifier la présence d'heteroscedasticité, nous avons estimé un second modèle, à l'aide eds écarts-types robustes de White (HC0 : utilisation justifiée par la taille conséquente de notre base de données)

6.2 Modèle 2 : MCO (ecarts-types robustes HC0)

```
Modèle 2: MCO, utilisant les observations 1-21602
Variable dépendante: l_price
Écarts-types robustes (hétéroscédasticité), variante HC0

                coefficient   p. critique
  ---------------------------------------
  const        -71,2690       1,13e-251  ***
  lat            1,55287      0,0000     ***
  long          -0,0774632    2,96e-06   ***
  sqft_living    0,000370797  0,0000     ***
  condition      0,0528027    2,41e-047  ***
  view           0,105866     4,96e-218  ***
  age            0,000443920  3,78e-06   ***

Moyenne var. dép.     13,04798   Éc. type var. dép.    0,526428
Somme carrés résidus  1879,648   Éc. type régression   0,295027

R2                    0,686004   R2 ajusté             0,685916

F(6, 21595)           4484,654   P. critique (F)       0,000000

Log de vraisemblance -4279,092   Critère d'Akaike      8572,185
Critère de Schwarz    8628,048   Hannan-Quinn          8590,394
```

Figure 10: Modèle 2

D'après les résultats d'estimations, tout les paramètres sont significatifs au seuil de 1%. Lors de l'analyse des relations et tendances des variables de notre base de données, nous avons détecté une possible relation non linéaire entre le prix et la surface habitable / l'age. Pour vérifier la significativité de l'ajout du carré de ces deux variables, un test de non-linéarité (carrés) a été réalisé.

```
Régression auxiliaire pour le test de non linéarité (termes au carré)
MCO, utilisant les observations 1-21602
Variable dépendante: uhat

                    coefficient        p. critique
  ------------------------------------------------
  sq_sqft_living      -2,31622e-08     1,60e-116  ***
  sq_age               6,66735e-05     1,22e-206  ***

  R2 non-ajusté = 0,129586

  Statistique de test: TR^2 = 2799,32,
  avec p. critique = P(Khi-deux(6) > 2799,32) = 0
```

Figure 11: Test de non-linéarité (carrés)

D'après les résultats du test, les deux variables au carré sont significatives au seuil de 1%. De plus, l'ajout de ces deux variables possède une réelle signification économique : en général, l'ajout d'un pied carré à un petit appartement n'aura pas le même effet que l'ajout d'un pied carré à une villa. Aussi, les logements peuvent atteindre un stade de vieillesse pour lequel l'ancienneté joue en faveur du prix (biens de luxe, biens anciens, de collection où à valeur historique) Nous avons donc décider d'estimer un troisième modèle en incluant ces deux variables.

6.3 Modèle 3 : MCO (ecarts-types robustes HC0) + termes au carré

```
Modèle 3: MCO, utilisant les observations 1-21602
Variable dépendante: l_price
Écarts-types robustes (hétéroscédasticité), variante HC0
```

	coefficient	p. critique		IC 99% low	IC 99% high
const	-68,4014	4,32e-232	***	-73,7522	-63,0506
lat	1,53622	0,0000	***	l1,50163	1,57082
long	-0,0600015	0,0003	***	-0,102778	-0,0172254
sqft_living	0,000483965	0,0000	***	0,000461537	0,000506392
condition	0,0686871	1,93e-081	***	0,0594692	0,0779050
view	0,114534	2,50e-267	***	0,106207	0,122862
age	-0,00693018	2,42e-171	***	-0,00756421	-0,00629616
sq_sqft_living	-2,22877e-08	3,64e-044	***	-2,63972e-08	-1,81782e-08
sq_age	7,34406e-05	2,17e-220	***	6,75399e-05	7,93413e-05

Moyenne var. dép.	13,04798	Éc. type var. dép.	0,526428
Somme carrés résidus	1750,386	Éc. type régression	0,284715
R2	0,707597	R2 ajusté	0,707489
F(8, 21593)	5643,912	P. critique (F)	0,000000
Log de vraisemblance	-3509,541	Critère d'Akaike	7037,081
Critère de Schwarz	7108,906	Hannan-Quinn	7060,493

Figure 12: Modèle 3

D'après les résultats d'estimations, tout les paramètres sont significatifs au seuil de 1%.
Les tests de Student joints montrent que l'effet joint des termes et de leur transformation au carré est significatif au seuil de 1%.
Aucun des intervalle de confiance ne contient 0, et aucun des écarts entre les bornes n'est éloigné de manière significative. Ainsi, les intervalles de confiance viennent consolider la confiance en les paramètre, et confirment les résultats des tests précédents au seuil de 1%.
Le R2 nous indique que 70.7597% de la variance est expliquée par le modèle, montrant une performance satisfaisante de notre modèle sans atteindre des valeurs extrêmes ($> 80\%$) pouvant être signe d'overfitting.
Le test de Fisher est significatif au seuil de 1% : le modèle est globalement

significatif pour expliquer le prix de vente des maisons.

En comparaison aux autres modèles, le modèle 3 est celui qui atteint le R2 ajusté le plus haut (0.707489). Ainsi, nous allons choisir ce modèle comme modèle final.
Selon ce modèle :
La constante (-68.4014) donne le log du prix lorsque toute les autres variables sont =0.
Le paramètres liés à Latitude (divisés par 10 pour correspondre au pas des graduations (0.1) de la carte du comté de King) montrent qu'en moyenne, le prix de vente des biens augmente de 15.36% lorsqu'on se déplace sur la carte, du sud au nord, de 0.1 unité (ceteris paribus)
Le paramètres liés à Longitude (divisés par 5 pour correspondre au pas des graduations (0.2) de la carte du comté de King) montrent qu'en moyenne, le prix de vente des biens baisse de -1.2% lorsqu'on se déplace sur la carte, d'ouest à est , de 0.2 unité (ceteris paribus)
L'ajout d'un pied carré permet de faire grimper le prix de vente de 0.048% (ceteris paribus) et cet effet est diminué de 0,000000159% par pieds carré ajouté. Ainsi, l'ajout d'un pied carré à un bien dont la superficie habitable est supérieure à 21714.4 pieds carré fait perdre de la valeur au bien en question. Les prix des biens baissent en moyenne de -0.69% par année d'ancienneté (ceteris paribus) et cet effet se réduit de 0.0073% par année d'ancienneté. Ainsi, les biens prennent de la valeur avec le temps à partir de 94 ans d'ancienneté.
En moyenne, un bien en meilleur état (dont la notation de l'état est supérieur de 1 à celle d'un autre bien) est vendu 6.87% plus cher qu'un bien concurrent.
En moyenne, un bien dont la vue est de meilleure qualité (dont la notation de la vue est supérieur de 1 à celle d'un autre bien) est vendu 11.45% plus cher qu'un bien concurrent.

Les signes de ces paramètres sont ceux attendus. Les coefficients sont cohérents avec la réalité.

7 Tests et discution

Afin de mesurer la performance et la robustesse de notre modèle sur la base de données, des tests concernant la distribution des résidus ainsi que des tests de ruptures structurelle ont été menés.

7.1 Colinéarité

```
Facteurs d'inflation de variance
Valeur minimale possible = 1.0
Des valeurs > 10.0 peuvent indiquer un problème de colinéarité

          lat    1,054
         long    1,258
   sqft_living    8,842
     condition    1,208
          view    1,140
           age   13,218
sq_sqft_living    8,321
        sq_age   12,090
```

Figure 13: Facteurs d'inflation de la variance

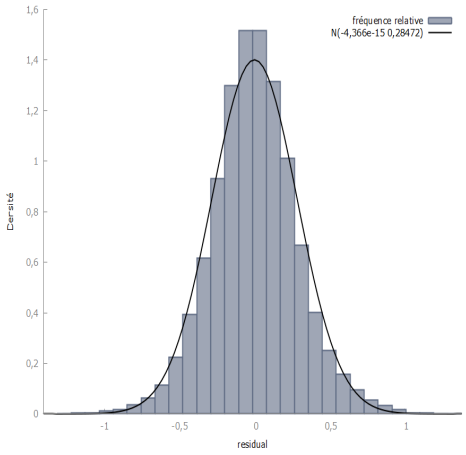
Selon ce test, aucune colinéarité critique n'est à constater pour notre modèle (les valeurs > 10 pour les variables age et sq_age sont justifiées par le fait que la variable age est naturellement colinéaire avec sa transformation au carré).

7.2 Normalité des résidus

```
Distribution des fréquences pour residual, obs 1-21602
nombre de classes = 29, moyenne -4,36598e-015, éc. type = 0,284715

Test de l'hypothèse nulle de normalité de la distribution:
Khi-deux(2) = 441,726 avec p. critique 0,00000
```

(a) Test de normalité des résidus



(b) Graphique issu du test

Figure 14: Test de normalité des résidus et visualisation graphique

Du point de vue analytique, le test de normalité des résidus indique que les résidus ne suivent pas exactement une loi normale (p. critique = 0,00000). Cependant, le graphique de distribution des résidus montre une distribution similaire à la loi normale (distribution centrée et réduite). La non-normalité des résidus peut être expliquée ici par la présence d'heteroscedasticité. Bien que corrigée en partie par les écarts-types de white, elle ne peut disparaitre complètement.

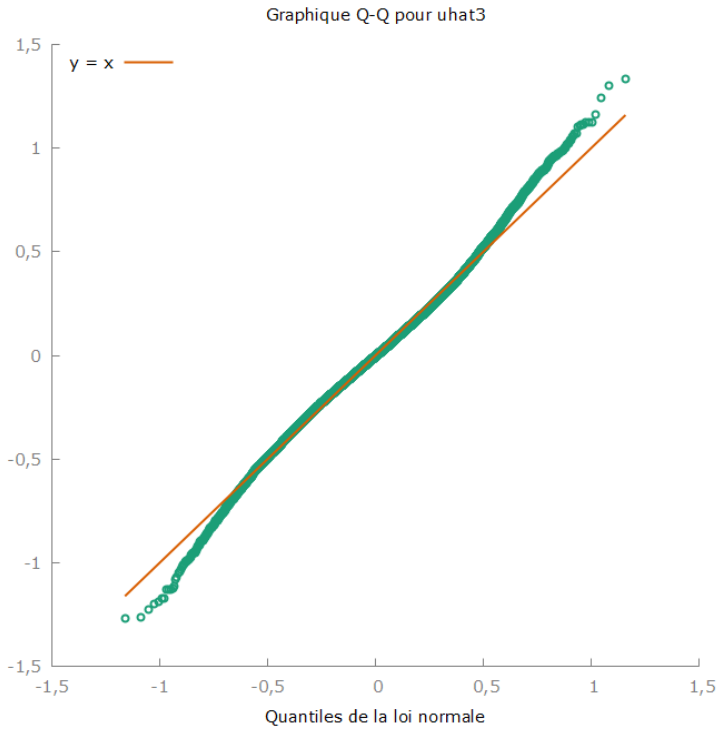


Figure 15: Graphique Q-Q des résidus

Le graphique Q-Q des résidus vient consolider les conclusions précédentes : les résidus se rapprochent fortement d'une distribution normale centrée réduite. En conclusion, malgré la présence d'hétéroscedasticité, le caractère réduit et centré de la distribution des erreurs indique que l'on peut porter une confiance significative en nos résultats d'estimation.

7.3 Observations influentes

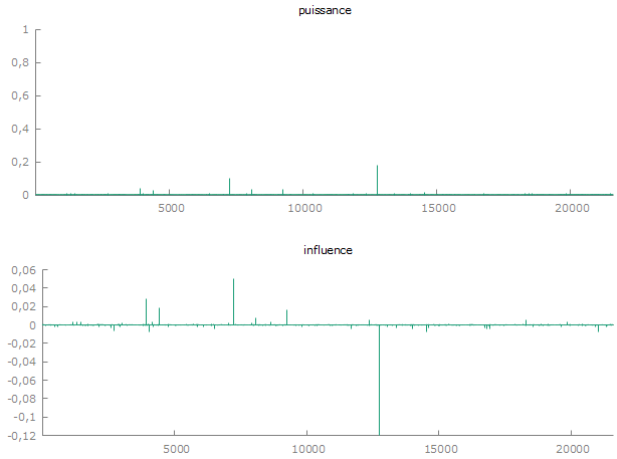


Figure 16: Levier et facteurs d'influence

A partir du graphique, nous pouvons identifier 1 observation influente (n=12770, dont le DFFITS (valeur absolue) = 0.995 significativement supérieur à $-2 * \sqrt{\frac{p}{n}} = 0.04$). Celle-ci n'étant pas issue d'une erreur de saisie dans la base de données, il est pertinent de la garder dans la dataframe car cette dernière a une influence significative sur le modèle.

7.4 Test de Chow (échantillon scindé en 2)

```
Régression augmentée pour le test de Chow
MCO, utilisant les observations 1-21602
Variable dépendante: l_price
Écarts-types robustes (hétéroscédasticité), variante HC0

Test de Chow pour rupture structurelle à l'observation 10801
  Khi-deux(9) = 17,0185 avec p. critique 0,0484
  F-norm: F(9, 21584) = 1,89095 avec p. critique 0,0485
```

Figure 17: Test de Chow (échantillon scindé en 2 sous-échantillons

Le test de Chow montre que lorsqu'on scinde la base de données en 2 sous-échantillons, aucune rupture structurelle significative n'est observée au seuil de 5%. Ainsi, ce test indique une certaine stabilité et similarité des paramètres estimés du modèle entre les deux sous-ensembles de données.

7.5 Test de Chow (renovated)

```
Régression augmentée pour le test de Chow
MCO, utilisant les observations 1-21602
Variable dépendante: l_price
Écarts-types robustes (hétéroscédasticité), variante HC0

                      coefficient      éc. type    t de Student  p. critique
  ------------------------------------------------------------------------------
  const               -66,7769         2,06053       -32,41       5,21e-225  ***
  lat                   1,54034        0,0134321     114,7        0,0000     ***
  long                 -0,0450629      0,0163162      -2,762      0,0058     ***
  sqft_living           0,000479763    8,93801e-06     53,68       0,0000     ***
  condition             0,0765584      0,00366583      20,88       6,61e-096  ***
  view                  0,112727       0,00334504      33,70       1,11e-242  ***
  age                  -0,00758890     0,000247710    -30,64       8,34e-202  ***
  sq_sqft_living       -2,26365e-08    1,66299e-09    -13,61       5,07e-042  ***
  sq_age                7,81084e-05    2,33728e-06     33,42       8,82e-239  ***
  renovated           -33,4520        13,1894          -2,536      0,0112     **
  re_lat               -0,0707241      0,105038        -0,6733     0,5008
  re_long              -0,305140       0,118846        -2,568      0,0102     **
  re_sqft_living        2,49212e-05    2,75110e-05      0,9059     0,3650
  re_condition         -0,0339730      0,0234018       -1,452      0,1466
  re_view              -0,00741563     0,0116672       -0,6356     0,5250
  re_age               -0,00481106     0,00349661      -1,376      0,1689
  re_sq_sqft_living     8,64114e-010   3,45427e-09      0,2502     0,8025
  re_sq_age             7,17706e-06    2,30505e-05      0,3114     0,7555

Test de Chow pour différence structurelle par rapport à renovated
  Khi-deux(9) = 160,949 avec p. critique 0,0000
  F-norm: F(9, 21584) = 17,8832 avec p. critique 0,0000
```

Figure 18: Test de Chow (renovated)

Le test de Chow pour la variable renovated montre une rupture structurelle significative entre les biens rénovés et non-rénovés. Seul le terme d'interaction entre renvated et long est significatif (à 5%) : lorsque l'on se déplace d'ouest à est, le prix des biens rénovés baisse davantage que celui des biens non-rénovés

7.6 Test de Chow (centre_ville)

```
Régression augmentée pour le test de Chow
MCO, utilisant les observations 1-21602
Variable dépendante: l_price
Écarts-types robustes (hétéroscédasticité), variante HC0

                      coefficient     éc. type    t de Student  p. critique
  ------------------------------------------------------------------------
  const               -38,8494        3,02139      -12,86       1,07e-037  ***
  lat                   1,37406       0,0151626     90,62        0,0000     ***
  long                  0,119008      0,0227996      5,220       1,81e-07   ***
  sqft_living           0,000470687   1,19146e-05   39,50        0,0000     ***
  condition             0,0628722     0,00509045    12,35        6,31e-035  ***
  view                  0,124161      0,00593580    20,92        3,36e-096  ***
  age                  -0,00342801    0,000369132   -9,287       1,74e-020  ***
  sq_sqft_living       -2,22723e-08   2,10641e-09   -10,57       4,58e-026  ***
  sq_age                1,97082e-05   4,35925e-06     4,521      6,19e-06   ***
  centre_ville         45,9620        6,07822         7,562      4,14e-014  ***
  ce_lat                0,259316      0,0381158       6,803      1,05e-011  ***
  ce_long               0,475734      0,0476022       9,994      1,82e-023  ***
  ce_sqft_living        2,38618e-05   1,67269e-05     1,427      0,1537
  ce_condition          0,00600768    0,00704849      0,8523     0,3940
  ce_view              -0,0181450     0,00713850     -2,542      0,0110     **
  ce_age               -0,00570590    0,000512509   -11,13       1,03e-028  ***
  ce_sq_sqft_living    -2,36317e-010  3,03663e-09    -0,07782    0,9380
  ce_sq_age             7,27646e-05   5,39927e-06    13,48       3,15e-041  ***

Test de Chow pour différence structurelle par rapport à centre_ville
  Khi-deux(9) = 794,483 avec p. critique 0,0000
  F-norm: F(9, 21584) = 88,2758 avec p. critique 0,0000
```

Figure 19: Test de Chow (centre_ville)

Le test de Chow pour la variable centre_ville montre une rupture structurelle significative entre les biens situés dans les villes de Seattle et Bellevue (selon leur code postal) des biens situés en zone rurale. Les termes d'interaction entre centre_ville et lat, long, view, age et sq_age sont significatifs au seuil de 1% (excepté view : 5%). L'influence de la géolocalisation sur le prix de vente est supérieure pour les biens situés au centre-ville. L'impact de l'ancienneté du logement ainsi que la baisse de son impact dans le temps sont plus fort sur la baisse des logements situés au centre_ville et l'impact de la qualité de la vue est moindre pour ces derniers.

7.7 Test de Chow (waterfront)

```
Régression augmentée pour le test de Chow
MCO, utilisant les observations 1-21602
Variable dépendante: l_price
Écarts-types robustes (hétéroscédasticité), variante HC0

                     coefficient     éc. type    t de Student  p. critique
  ------------------------------------------------------------------------
  const              -68,7097        2,06747       -33,23       3,10e-236  ***
  lat                  1,53610       0,0133901     114,7        0,0000     ***
  long                -0,0625405     0,0165019      -3,790      0,0002     ***
  sqft_living          0,000488529   8,66283e-06     56,39       0,0000     ***
  condition            0,0686305     0,00357926      19,17       2,87e-081  ***
  view                 0,0974699     0,00325559      29,94       5,31e-193  ***
  age                 -0,00696329    0,000245328    -28,38       5,03e-174  ***
  sq_sqft_living      -2,30783e-08   1,60179e-09    -14,41       7,62e-047  ***
  sq_age               7,39314e-05   2,28495e-06     32,36       2,58e-224  ***
  waterfront         106,793         34,3045          3,113      0,0019     ***
  wa_lat               0,218769      0,255730        0,8555     0,3923
  wa_long              0,952058      0,212544        4,479      7,53e-06   ***
  wa_sqft_living      -4,46497e-05   5,09511e-05    -0,8763     0,3809
  wa_condition         0,0120119     0,0329801       0,3642     0,7157
  wa_view             -0,0928299     0,0377629      -2,458      0,0140     **
  wa_age               0,00459183    0,00311413      1,475      0,1404
  wa_sq_sqft_living    1,98217e-09   5,19958e-09     0,3812     0,7030
  wa_sq_age           -6,02442e-05   2,88838e-05    -2,086      0,0370     **

Test de Chow pour différence structurelle par rapport à waterfront
  Khi-deux(9) = 424,952 avec p. critique 0,0000
  F-norm: F(9, 21584) = 47,2169 avec p. critique 0,0000
```

Figure 20: Test de Chow (waterfront)

Le test de Chow pour la variable waterfront montre une rupture structurelle significative entre les biens situés en zone côtière et les autres biens. Les termes d'interaction entre waterfront et long, view, sq_age sont significatifs à 5% (excepté long : 1%). Les biens en zone côtière subissent une augmentation du prix supérieure aux autres biens lorsqu'on se déplace du sud au nord. l'effet de la qualité de la vue et l'atténuation de l'effet de l'age sur le prix est moindre pour les biens situés en zone côtière.

8 Conclusion

Le projet visait à analyser les principaux déterminent du prix des biens immobiliers dans le comté de King, en utilisant la méthode des MCO sur la base de données.
Les résultats ont montré des corrélations significatives entre le prix des biens et divers facteurs comme la superficie habitable, la qualité de la vue, l'état du bien et la localisation géographique. Des modèles non-linéaires ont également été explorés pour une meilleure précision.
Les résultats des tests menés permettent de constituer un modèle significatif et robuste pour expliquer le prix des biens vendus dans le comté de King, avec plusieurs variables explicatives offrant une description quantifiée des différents élément influençant le prix de vente.
Le projet reconnaît certaines limites, comme la mesure des spécificités géographiques des résultats ou les ruptures structurelles constatées en fonction de certaines variables dummy, suggérant des recherches futures dans les déterminents des variables centre_ville, renovated ou waterfront pour permettre une précision accrue des modèles, tout en étant attentif face au risque d'overfitting.

Le projet a donc réussi à identifier et quantifier les principaux facteurs influençant les prix immobiliers dans le comté de King, offrant une contribution significative à la compréhension du marché immobilier dans cette région.