

# E2-L3S2 Le prix des logements dans l'État de Washington aux États-Unis

Lola LUBIN et Alexis VIALATTE

2021

## Table des matières

<b>1 Formalisation</b>	<b>2</b>
<b>2 La base de données BDD_data.csv</b>	<b>3</b>
2.1 Les manipulations de la base de données . . . . .	4
2.1.1 Les variables <code>id</code> et <code>date</code> . . . . .	4
2.1.2 Les variables <code>sqft_living</code> et <code>sqft_lot</code> . . . . .	5
2.1.3 Les variables <code>bedrooms</code> , <code>bathrooms</code> et <code>floors</code> . . . . .	6
2.1.4 Les variables <code>sqft_above</code> et <code>sqft_basement</code> . . . . .	7
2.1.5 Les variables <code>yr_built</code> et <code>yr_renovated</code> . . . . .	8
2.1.6 Les variables <code>waterfront</code> , <code>view</code> , <code>condition</code> et <code>grade</code> . . . . .	9
2.1.7 Les variables <code>zipcode</code> , <code>lat</code> et <code>long</code> . . . . .	10
<b>3 La nouvelle base de données BDD_data.csv</b>	<b>12</b>
<b>4 <math>\mathcal{A}_1</math> : l'analyse du prix d'un logement en fonction de sa superficie intérieure en pieds carrés et de sa superficie extérieure en pieds carrés</b>	<b>13</b>
4.1 Analyse des résultats . . . . .	14
4.2 Analyse des résidus . . . . .	15
4.3 Analyse du modèle 5 moins les résidus aberrants pour la distance de Cook . . . . .	17
4.3.1 Analyse des résultats . . . . .	17
4.3.2 Analyse des résidus . . . . .	18
4.4 $\mathcal{A}_1^H$ : l'analyse de l'hétérosécédasticité . . . . .	20
<b>5 <math>\mathcal{A}_2</math> : l'analyse du prix d'un logement en fonction de son nombre de chambres, de son nombre de salles de bain et de son nombre d'étages</b>	<b>23</b>
5.1 Analyse des résultats . . . . .	24
5.2 Analyse des résidus . . . . .	25
5.3 $\mathcal{A}_2^H$ : l'analyse de l'autocorrélation . . . . .	27
<b>6 <math>\mathcal{A}_3</math> : l'analyse du prix d'un logement à partir de 645 000 dollars en fonction de sa vue sur la mer ou non, de son point de vue et de son <i>design</i></b>	<b>31</b>
6.1 Analyse des résultats . . . . .	33
6.2 Analyse des résidus . . . . .	34
<b>7 <math>\mathcal{A}_4</math> : l'analyse du prix d'un logement en fonction de son état</b>	<b>36</b>
7.1 Analyse des résultats . . . . .	37
7.2 Analyse des résidus . . . . .	38
7.3 $\mathcal{A}_4^H$ : l'analyse de l'autocorrélation . . . . .	40
<b>8 <math>\mathcal{A}_5</math> : l'analyse du prix d'un logement en fonction de son âge de construction et de son âge de rénovation</b>	<b>44</b>
8.1 Analyse des résultats . . . . .	45

8.2 Analyse des résidus . . . . .	46
<b>9 <math>\mathcal{A}_6</math> : le prix d'un logement dans la ville de Seattle dans l'État de Washington aux États-Unis</b>	<b>48</b>
<b>10 Synthèse</b>	<b>48</b>
<b>11 tools</b>	<b>48</b>
<b>12 packages</b>	<b>51</b>

# 1 Formalisation

---

Nous posons  $x_{i,k}$  le  $i$ -ième logement et la  $k$ -ième variable avec  $i \in [1, \dots, I]$  et  $k \in [1, \dots, K]$  ainsi que  $\mathcal{M}_j$  le  $j$ -ième modèle avec  $j \in [1, \dots, J]$  et  $\mathcal{A}_n$  la  $n$ -ième analyse avec  $n \in [1, \dots, N]$ . Nous aurons  $\forall j \in [1, \dots, J]$  et  $\forall n \in [1, \dots, N]$  :

$$\mathcal{A}_n = \{\mathcal{M}_1, \dots, \mathcal{M}_J\}$$

Nous poserons la variable expliquée  $y_i = x_{i,\tilde{k}}$  où  $\tilde{k} \in [1, \dots, K]$ , les variables explicatives  $x_{i,k}$  où  $k \in [1, \dots, K] \setminus \tilde{k}$  et la variable résiduelle  $z_i$ . Nous aurons :

$$\mathcal{M}_j : y_i = \alpha + \sum_{k=1}^K (\beta_k x_{i,k}) + z_i \quad (1)$$

Nous minimiserons la somme des résidus au carré. Nous aurons :

$$\mathcal{M}_j : \min \left( \sum_{i=1}^I z_i^2 \right) = \min \left\{ \sum_{i=1}^I \left[ y_i - \alpha - \sum_{k=1}^K (\beta_k x_{i,k}) \right]^2 \right\} \quad (2)$$

Nous chercherons à estimer  $\hat{\alpha}$ , la constante du  $j$ -ième modèle et  $\hat{\beta}_k$ , le coefficient de la  $k$ -ième variable du  $j$ -ième modèle.

Nous poserons  $\phi$  le risque de première espèce,  $t$  la statistique du test de Students,  $F$  la statistique du test de Fisher,  $DW$  la statistique du test de Durbin-Watson,  $W$  la statistique du test de Shapiro-Wilk,  $BP$  la statistique du test de Breusch-Pagan et  $HMC$  la statistique du test d'Harrison-McCabe. Nous aurons :

$$\begin{aligned} \mathcal{H}_0 : \alpha = 0 &\iff p\text{-value} \geq \phi \\ \mathcal{H}_1 : \alpha \neq 0 &\iff p\text{-value} < \phi \end{aligned}$$

$$\begin{aligned} \mathcal{H}_0 : \beta_k = 0 &\iff p\text{-value} \geq \phi \\ \mathcal{H}_1 : \beta_k \neq 0 &\iff p\text{-value} < \phi \end{aligned}$$

$$\begin{aligned} \mathcal{H}_0 : \text{absence de significativité} &\iff p\text{-value} \geq \phi \\ \mathcal{H}_1 : \text{significativité} &\iff p\text{-value} < \phi \end{aligned}$$

$$\begin{aligned} \mathcal{H}_0 : \text{absence d'autocorrélation} &\iff p\text{-value} \geq \phi \\ \mathcal{H}_1 : \text{autocorrélation} &\iff p\text{-value} < \phi \end{aligned}$$

$$\begin{aligned} \mathcal{H}_0 : \text{distribution normale} &\iff p\text{-value} \geq \phi \\ \mathcal{H}_1 : \text{absence de distribution normale} &\iff p\text{-value} < \phi \end{aligned}$$

$$\begin{aligned} \mathcal{H}_0 : \text{homoscédasticité} &\iff p\text{-value} \geq \phi \\ \mathcal{H}_1 : \text{hétéroscédisticité} &\iff p\text{-value} < \phi \end{aligned}$$


---

## 2 La base de données BDD\_data.csv

$$\forall n \in [1, \dots, N], \forall j \in [1, \dots, J] \iff \tilde{k} = 3$$

Nous allons commencer par nous poser une question : quelles sont les variables qui expliquent le prix des logements dans l'État de Washington aux États-Unis ?

TABLE 1 – Résumé de la variable price :

Moyenne	Ecart-type	Minimum	Q1	Q2	Q3	Maximum
540,088.100	367,127.200	75,000	321,950	450,000	645,000	7,700,000

Pour répondre à cette question, nous disposons d'une base de données `BDD_data.csv` contenant 21613 logements et 21 variables. Il y a 0 donnée manquante concernant 0 logement. Les variables sont :

1. id,
2. date,
3. price,
4. bedrooms,
5. bathrooms,
6. sqft\_living,
7. sqft\_lot,
8. floors,
9. waterfront,
10. view,
11. condition,
12. grade,
13. sqft\_above,
14. sqft\_basement,
15. yr\_built,
16. yr\_renovated,
17. zipcode,
18. lat,
19. long,
20. sqft\_living15,
21. sqft\_lot15.

Pour chaque logement, la variable `id` correspond à son index, la variable `date` correspond à son année de vente, la variable `price` correspond à son prix de vente, la variable `bedrooms` correspond à son nombre de chambres, la variable `bathrooms` correspond à son nombre de salles de bain, la variable `sqft_living` correspond à sa surface intérieure en pieds carrés, la variable `sqft_lot` correspond à sa surface totale en pieds carrés, la variable `floors` correspond à son nombre d'étages, la variable `waterfront` répond à s'il a une vue sur la mer ou non, la variable `view` correspond à son point de vue, la variable `condition` correspond à son état, la variable `grade` correspond à son *design*, la variable `sqft_above` correspond à sa surface intérieure en pieds carrés moins son sous-sol, la variable `sqft_basement` correspond à sa surface intérieure en pieds carrés dans son sous-sol, la variable `yr_built` correspond à son année de construction, la variable `yr_renovated` correspond à son année de rénovation s'il a été rénové, la variable `zipcode` correspond à son code postal, les variables `lat` et `long` correspondent à ses coordonnées géographiques et les variables `sqft_living15` et `sqft_lot15` correspondent à sa surface intérieure et à sa surface totale en 2015.

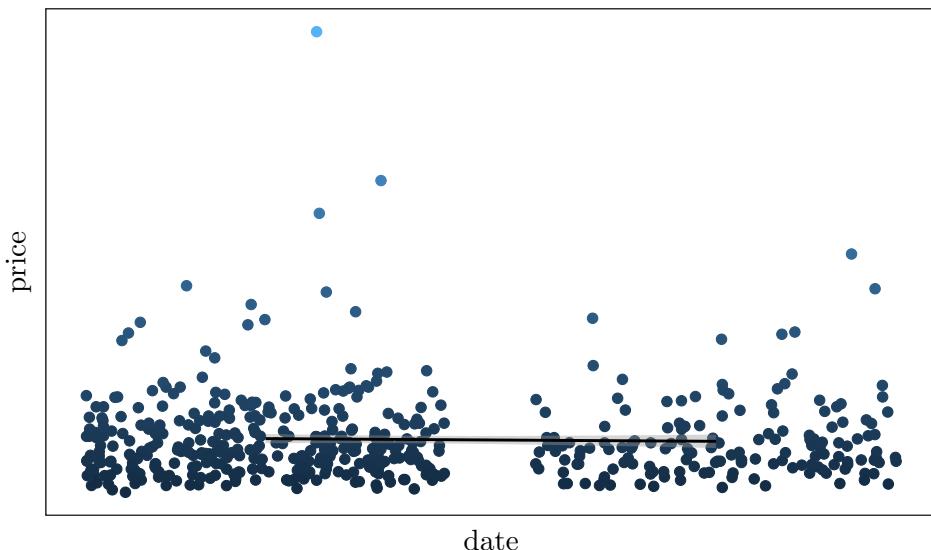
## 2.1 Les manipulations de la base de données

Nous avons discuté de la base de données et nous ne pensons pas conserver toutes les variables principalement pour des raisons techniques. Nous allons regarder chaque variable de la base de données et nous allons choisir de la conserver, de la transformer ou de la supprimer.

### 2.1.1 Les variables `id` et `date`

```
data$date <- as.character( data$date )  
  
data$date <- str_sub( data$date,  
                      1,  
                      4 )  
  
data$date <- as.numeric( data$date )
```

La variable `id` n'est pas intéressante à analyser. Nous nous sommes demandés si, pour chaque logement, la date de la vente avait une influence sur le prix de ces derniers. *A priori*, nous supposons que non. Nous ne nous servirons donc pas des variables `id` et `date` car le graphique ci-dessous ne va pas à l'encontre de notre avis.

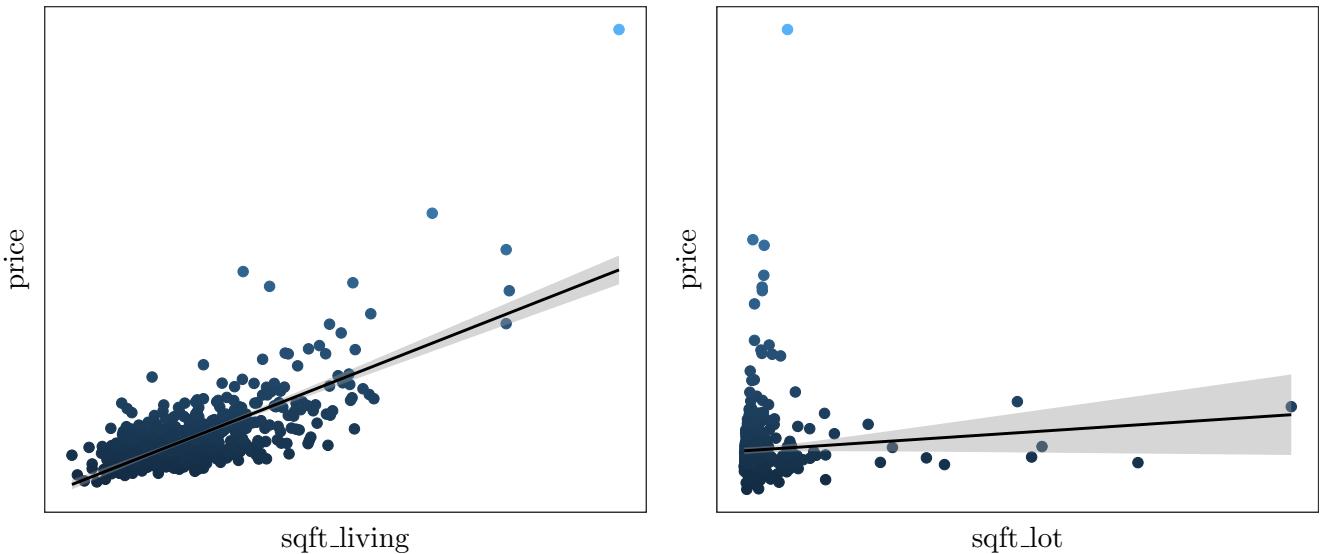


### 2.1.2 Les variables `sqft_living` et `sqft_lot`

TABLE 2 – Résumé des variables price, sqft-living et sqft-garden :

	price	sqft_living	sqft_lot
Moyenne	540088.1	2079.900	15106.97
Ecart-type	367127.2	918.441	41420.51
Minimum	75000.0	290.000	520.00
Q1	321950.0	1427.000	5040.00
Q2	450000.0	1910.000	7618.00
Q3	645000.0	2550.000	10688.00
Maximum	7700000.0	13540.000	1651359.00

Nous nous sommes demandés si, pour chaque logement, la surface intérieure en pieds carrés avait une influence plus ou moins importante que la surface extérieure en pieds carrés sur le prix de ces derniers. *A priori*, nous supposons que oui car un pied carré intérieur est plus important qu'un pied carré extérieur. Nous nous servirons donc des variables `sqft_living` et `sqft_lot` car les graphiques ci-dessous ne vont pas à l'encontre de notre avis.

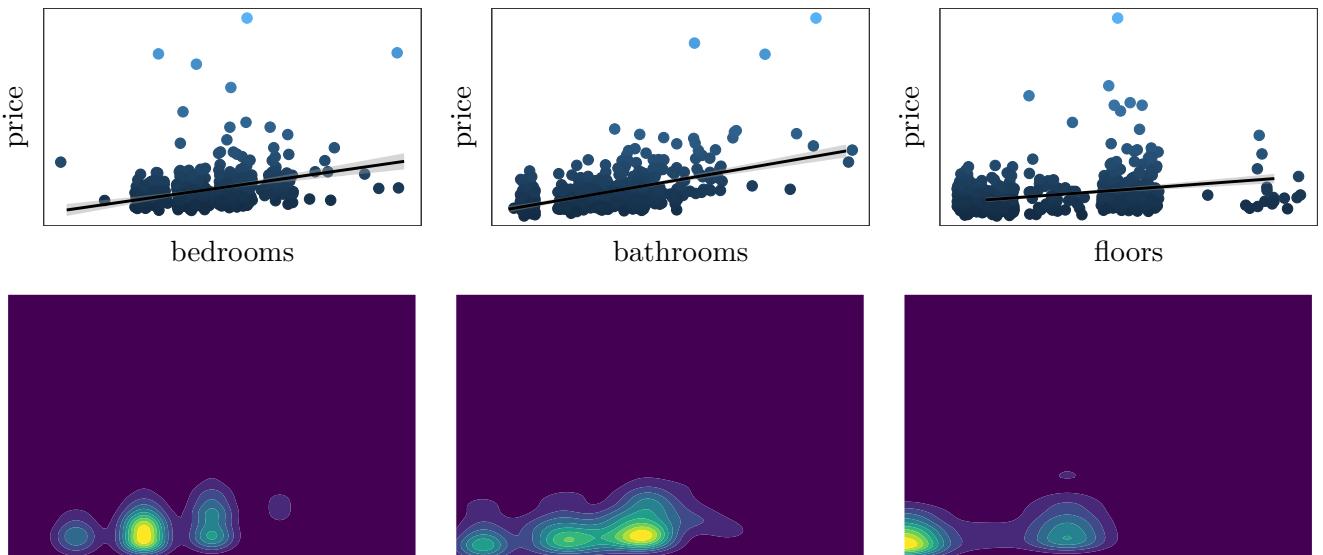


### 2.1.3 Les variables bedrooms, bathrooms et floors

TABLE 3 – Résumé des variables bedrooms, bathrooms et floors :

	bedrooms	bathrooms	floors
Moyenne	3.371	2.115	1.494
Ecart-type	0.930	0.770	0.540
Minimum	0.000	0.000	1.000
Q1	3.000	1.750	1.000
Q2	3.000	2.250	1.500
Q3	4.000	2.500	2.000
Maximum	33.000	8.000	3.500

Nous nous sommes demandés si, pour chaque logement, le nombre de chambres, le nombre de salles de bain et le nombre d'étages avaient une influence sur le prix de ces derniers. *A priori*, nous supposons que oui car plus le nombre de chambres, le nombre de salles de bain et le nombre d'étages sont élevés et plus la surface intérieure en pieds carrés est élevée. Nous nous servirons donc des variables `bedrooms`, `bathrooms` et `floors` car les graphiques ci-dessous ne vont pas à l'encontre de notre avis.



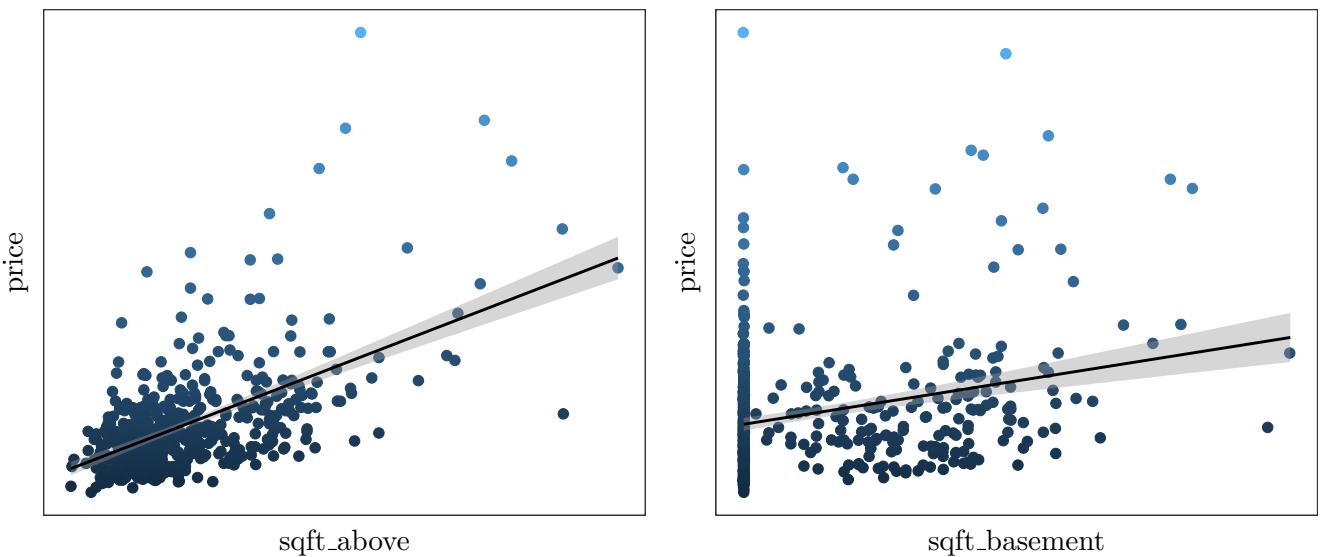
#### 2.1.4 Les variables `sqft_above` et `sqft_basement`

Les variables `sqft_built`, `sqft_renovated`, `sqft_living15` et `sqft_lot15` ne sont pas intéressantes à analyser car elles sont équivalentes des variables `yr_built` et `yr_renovated`.

TABLE 4 – Résumé des variables sqft-above et sqft-basement :

	sqft_above	sqft_basement
Moyenne	1788.391	291.509
Ecart-type	828.091	442.575
Minimum	290.000	0.000
Q1	1190.000	0.000
Q2	1560.000	0.000
Q3	2210.000	560.000
Maximum	9410.000	4820.000

Nous nous sommes demandés si, pour chaque logement, la surface intérieure en pieds carrés moins son sous-sol avait une influence plus ou moins importante que la surface intérieure en pieds carrés dans son sous-sol sur le prix de ces derniers. *A priori*, nous supposons que oui car un pied carré intérieur est plus important qu'un pied carré intérieur dans son sous-sol. Les graphiques ci-dessous ne vont pas à l'encontre de notre avis. Néanmoins, nous ne nous servirons pas des variables `sqft_above` et `sqft_basement` car elles interagissent avec la variable `sqft_living`.



### 2.1.5 Les variables yr\_built et yr\_renovated

Les variables `yr_built` et `yr_renovated` ne sont pas intéressantes à analyser. Nous allons créer les variables `life_built` et `life_renovated` :

$$\forall i \in [1, \dots, I] \iff x_{i,23} = x_{i,2} - x_{i,15}$$

$$\forall i \in [1, \dots, I] \iff x_{i,24} = x_{i,2} - x_{i,16}$$

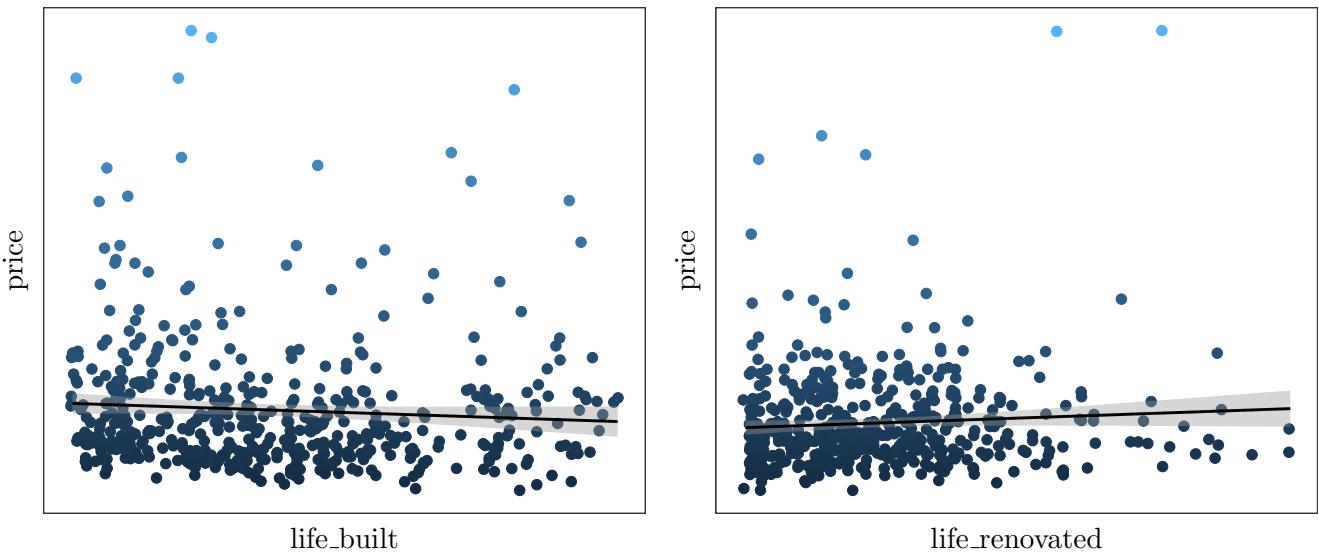
```
data$life_built <- data$date - data$yr_built[ data$yr_built != 0 ]
```

```
data$life_renovated <- data$date - data$yr_renovated[ data$yr_renovated != 0 ]
```

TABLE 5 – Résumé des variables life-built et life-renovated :

	life_built	life_renovated
Moyenne	43.318	18.492
Ecart-type	29.375	15.511
Minimum	-1.000	-1.000
Q1	18.000	7.000
Q2	40.000	15.000
Q3	63.000	27.000
Maximum	115.000	81.000

Nous nous sommes demandés si, pour chaque logement, l'âge de la construction et l'âge de la rénovation avaient une influence sur le prix de ces derniers. *A priori*, nous supposons que oui car plus le logement est récent et plus le prix de ce dernier est élevé. Nous nous servirons des variables `life_built` et `life_renovated` même si les graphiques ci-dessous vont à l'encontre de notre avis.



### 2.1.6 Les variables `waterfront`, `view`, `condition` et `grade`

Les variables `waterfront`, `view`, `condition` et `grade` sont des variables catégorielles :

```
data$waterfront <- as.factor( data$waterfront )

data$view <- as.factor( data$view )

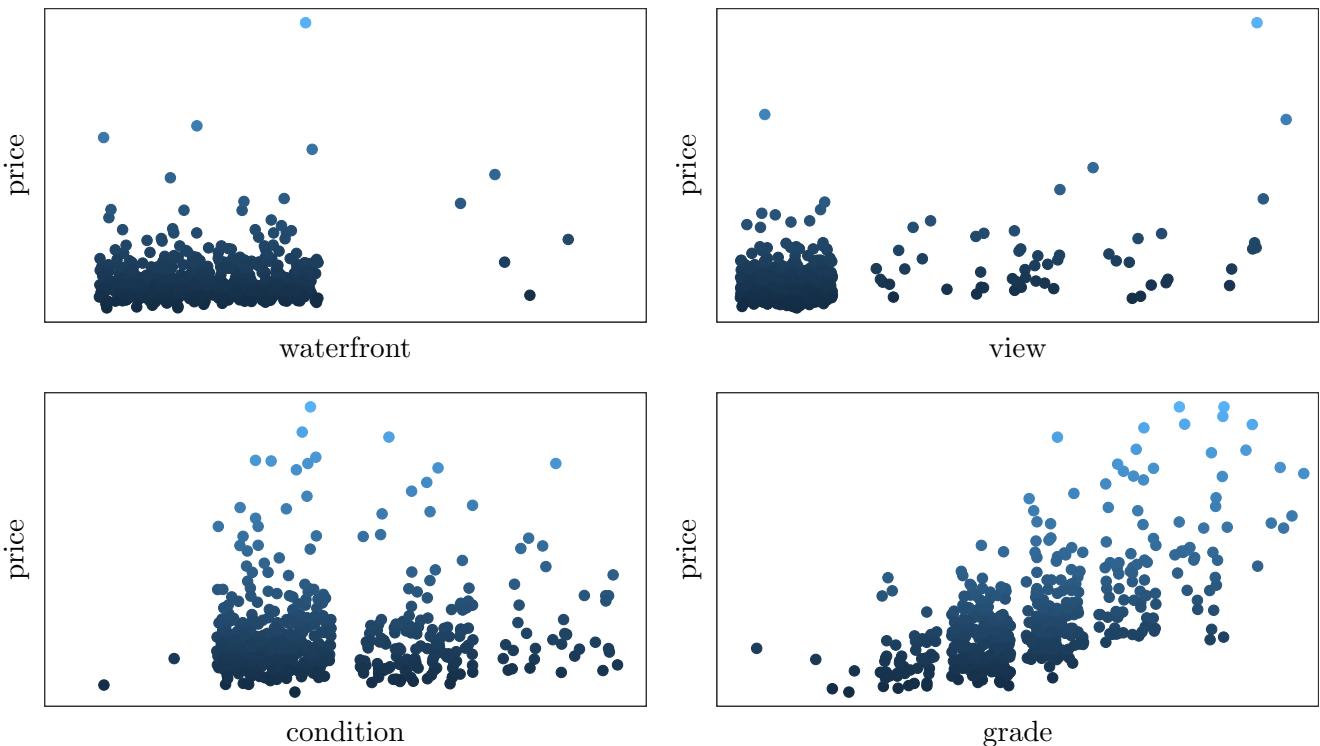
data$condition <- as.factor( data$condition )

data$grade <- as.factor( data$grade )
```

TABLE 6 – Résumé des variables `waterfront`, `view`, `condition` et `grade` :

	waterfront	view	condition	grade
Moyenne	1.008	1.234	3.409	6.657
Ecart-type	0.087	0.766	0.651	1.175
Minimum	1.000	1.000	1.000	1.000
Q1	1.000	1.000	3.000	6.000
Q2	1.000	1.000	3.000	6.000
Q3	1.000	1.000	4.000	7.000
Maximum	2.000	5.000	5.000	12.000

Nous nous sommes demandés si, pour chaque logement, la vue sur la mer ou non, le point de vue, l'état et le *design* avaient une influence sur le prix de ces derniers. *A priori*, nous supposons que oui car plus son point de vue est beau, plus l'état est bon, plus son *design* est classe et plus le prix de ce dernier est élevé. Nous nous servirons des variables `waterfront`, `view`, `condition` et `grade` même si les graphiques ci-dessous vont à l'encontre de notre avis.



### 2.1.7 Les variables zipcode, lat et long

Les variables `zipcode`, `lat` et `long` sont compliquées à analyser. Nous avons la variable `zipcode` qui est comprise entre 98001 et 98199 et les variables `lat` et `long` qui sont comprises, respectivement, entre 47.16 et 47.78 et -122.52 et -121.32. La base de données `BDD_data.csv` correspond grossièrement à la ville de Seattle dans l'État de Washington aux États-Unis.

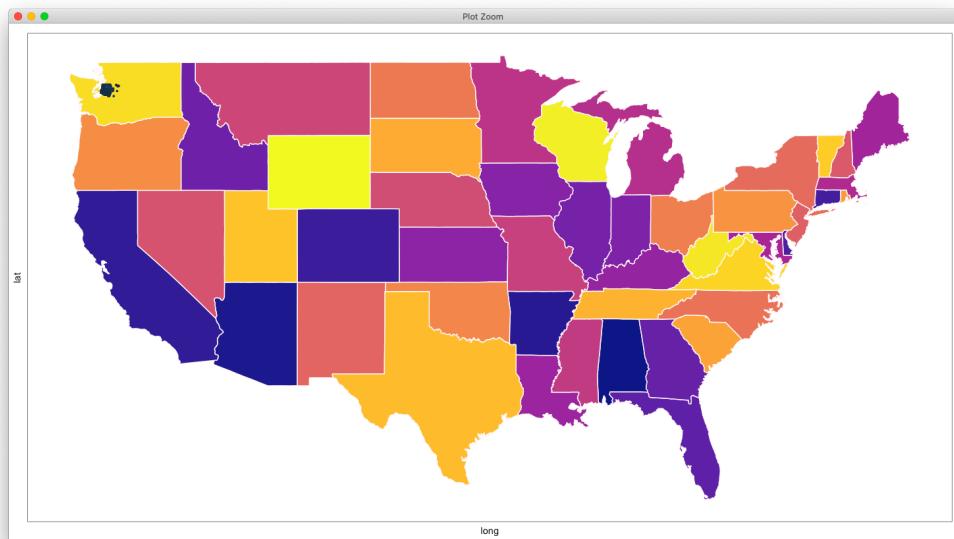


FIGURE 1 – Les États-Unis



FIGURE 2 – Zoom sur l'État de Washington

Est-ce qu'un logement à l'ouest coûte moins cher qu'un logement à l'est ou inversement ? Est-ce qu'un logement au nord coûte moins cher qu'un logement au sud ou inversement ? Nous n'avons aucun *a priori*.

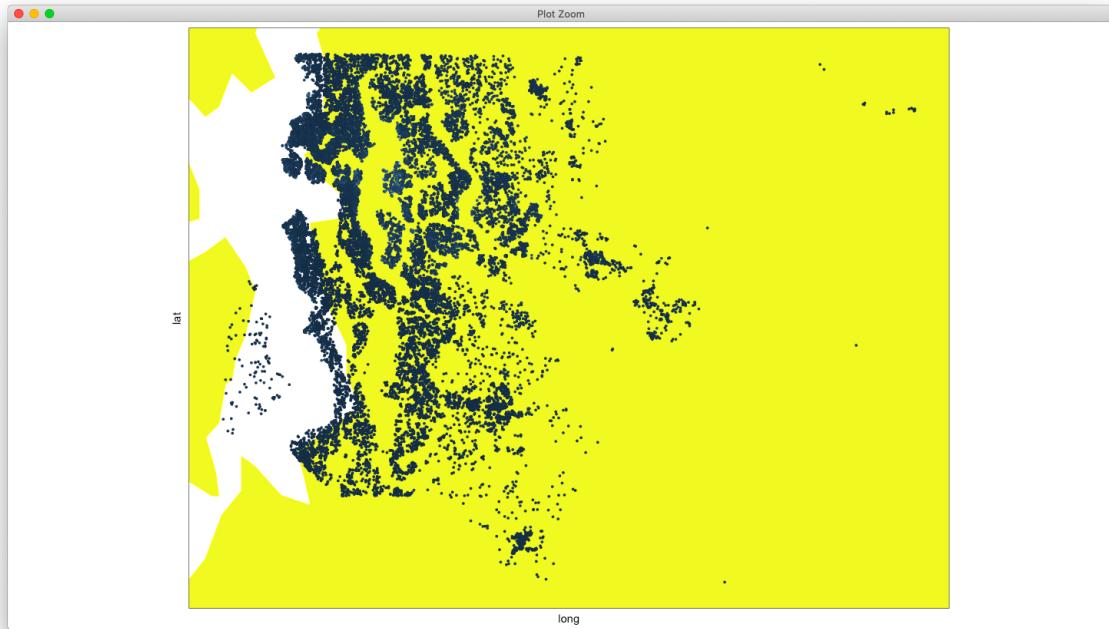


FIGURE 3 – Zoom sur le nuage de points

D'après la carte ci dessus, nous supposons que plus un logement est au centre de Seattle et plus le prix de ce dernier est élevé. Nous allons zoomer sur le nuage de points le plus clair.

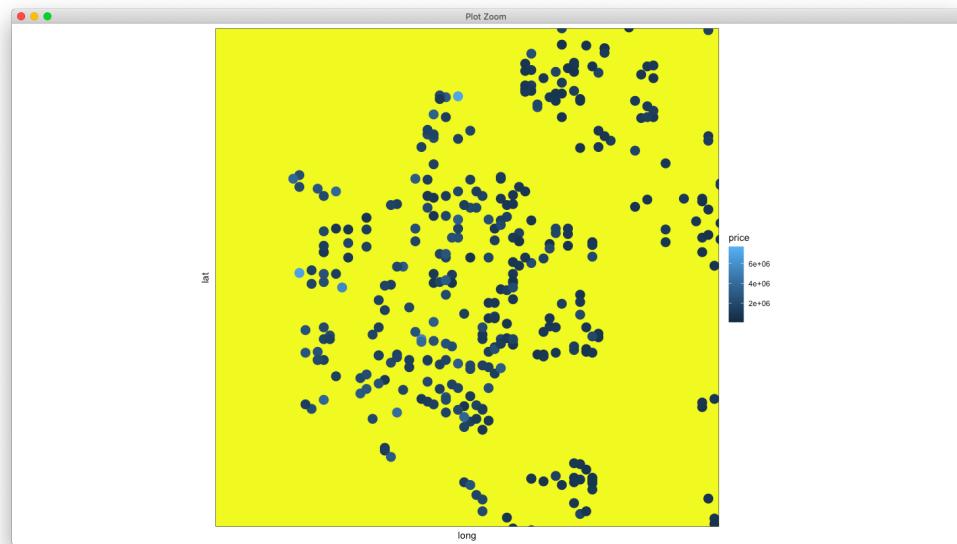


FIGURE 4 – Zoom sur le centre de Seattle

Nous ne nous servirons pas des variables `zipcode`, `lat` et `long` car elles ne sont pas intéressantes à régresser linéairement.

### 3 La nouvelle base de données BDD\_data.csv

TABLE 7 – Aperçu de la base de la nouvelle base de données :

	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	life_built	life_renovated
10609	367000	3	1.75	2000	12669	1 1		4 4		6	49	14
1204	510000	3	1.75	1490	3800	1 1		1 3		5	102	5
6262	848000	5	1.75	2290	4320	2 1		1 3		6	87	2
9621	473000	3	1.00	1280	10000	1 1		1 4		6	60	57
17361	840000	4	1.75	2330	4000	2 1		1 5		7	90	10
1400	408000	2	2.00	1200	3900	1 1		1 3		7	9	17
5724	96500	3	1.00	840	12091	1 1		1 3		5	55	0
18774	322500	3	2.00	1350	14200	1 1		1 3		6	25	50
3965	364950	4	2.50	2310	8030	2 1		1 3		6	36	34
9097	384000	6	3.00	2320	4502	1 1		1 4		6	27	19

TABLE 8 – Coefficients de corrélation linéaire de Pearson de la nouvelle base de données :

	price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	life_built	life_renovated
price	1.000	0.308	0.525	0.702	0.090	0.257	-0.054	-0.003
bedrooms	0.308	1.000	0.516	0.577	0.032	0.175	-0.154	-0.001
bathrooms	0.525	0.516	1.000	0.755	0.088	0.501	-0.506	-0.003
sqft_living	0.702	0.577	0.755	1.000	0.173	0.354	-0.318	-0.002
sqft_lot	0.090	0.032	0.088	0.173	1.000	-0.005	-0.053	0.008
floors	0.257	0.175	0.501	0.354	-0.005	1.000	-0.490	-0.009
life_built	-0.054	-0.154	-0.506	-0.318	-0.053	-0.490	1.000	-0.003
life_renovated	-0.003	-0.001	-0.003	-0.002	0.008	-0.009	-0.003	1.000

D'après la table 8, nous supposons que les coefficients adossés aux variables `bedrooms`, `bathrooms`, `sqft_living`, `sqft_lot` et `floors` sont strictement supérieurs à 0 et que les coefficients adossés aux variables `life_built` et `life_renovated` sont strictement inférieurs à 0.

#### 4 $\mathcal{A}_1$ : l'analyse du prix d'un logement en fonction de sa superficie intérieure en pieds carrés et de sa superficie extérieure en pieds carrés

$$\mathcal{A}_1 = \{\mathcal{M}_1, \mathcal{M}_3, \mathcal{M}_4, \mathcal{M}_4, \mathcal{M}_5, \mathcal{M}_6\}$$

TABLE 9 – Régressions linéaires de l'analyse 1 :

	<i>Dependent variable:</i>		
	price		
	(1)	(2)	(3)
sqft_living	282.879*** (1.964)	280.624*** (1.936)	
sqft_lot	-0.289*** (0.044)		0.795*** (0.060)
Constant	-43,900.230*** (4,398.565)	-43,580.740*** (4,402.690)	528,082.600*** (2,647.505)
Observations	21,613	21,613	21,613
R <sup>2</sup>	0.494	0.493	0.008
Adjusted R <sup>2</sup>	0.494	0.493	0.008
Residual Std. Error	261,192.300 (df = 21610)	261,452.900 (df = 21611)	365,657.000 (df = 21611)
F Statistic	10,543.990*** (df = 2; 21610)	21,001.910*** (df = 1; 21611)	175.140*** (df = 1; 21611)

Note:

\*p<0.1; \*\*p<0.05; \*\*\*p<0.01

TABLE 10 – Régressions linéaires de l'analyse 1 :

	<i>Dependent variable:</i>		
	log(price)		
	(1)	(2)	(3)
log(sqft_living)	0.873*** (0.007)	0.837*** (0.006)	
log(sqft_lot)	-0.053*** (0.003)		0.080*** (0.004)
Constant	6.930*** (0.048)	6.730*** (0.047)	12.325*** (0.036)
Observations	21,613	21,613	21,613
R <sup>2</sup>	0.463	0.456	0.019
Adjusted R <sup>2</sup>	0.463	0.455	0.019
Residual Std. Error	0.386 (df = 21610)	0.389 (df = 21611)	0.522 (df = 21611)
F Statistic	9,310.386*** (df = 2; 21610)	18,079.140*** (df = 1; 21611)	417.860*** (df = 1; 21611)

Note:

\*p<0.1; \*\*p<0.05; \*\*\*p<0.01

## 4.1 Analyse des résultats

TABLE 11 – Aperçu de la base de données de l'analyse 1 :

	price	sqft_living	sqft_lot
4789	716100	1640	4000
13910	695000	2650	9990
8533	350000	900	6380
11667	650000	1910	16532
12078	285000	1930	7200
18904	435000	670	1800
16764	829000	2690	10443
12123	251000	840	4495
19044	872000	2860	40284
8676	242000	2340	7494

TABLE 12 – Comparaison des performances des régressions linéaires de l'analyse 1

Name	Model	AIC	BIC	R2	R2_adjusted	RMSE	Sigma	Performance_Score
model4	lm	20195.57	20227.50	0.463	0.463	0.386	0.386	0.979
model5	lm	20486.98	20510.93	0.456	0.455	0.389	0.389	0.974
model6	lm	33211.62	33235.57	0.019	0.019	0.522	0.522	0.667
model1	lm	600498.46	600530.39	0.494	0.494	261174.188	261192.316	0.437
model2	lm	600540.57	600564.51	0.493	0.493	261440.790	261452.888	0.436
model3	lm	615040.37	615064.31	0.008	0.008	365640.079	365656.998	0.000

Nous allons analyser le modèle 5.

$$\mathcal{M}_5 : \log(y_i) = \alpha + \beta_6 \log(x_{i,6}) + z_i \iff \mathcal{M}_5 : \log(y_i) = 6,730 + 0,837 \log(x_{i,6}) + z_i$$

Pour  $\alpha$ , nous avons moins de 0,1% de chance de se tromper en rejetant l'hypothèse nulle. Par conséquent, nous rejetons l'hypothèse nulle et nous en déduisons que la constante du modèle 5 est différente de 0 au seuil de 0,1%.

Pour  $\beta_6$ , nous avons moins de 0,1% de chance de se tromper en rejetant l'hypothèse nulle. Par conséquent, nous rejetons l'hypothèse nulle et nous en déduisons que le coefficient du modèle 5 de la variable explicative `sqft_living` est différent de 0 au seuil de 0,1%.

Nous avons  $R^2 = 0,456$  dans le modèle 5 donc 45,6% de la variation du prix des logements est expliquée par la surface intérieure de ces derniers.

Pour  $F$ , nous avons moins de 0,1% de chance de se tromper en rejetant l'hypothèse d'absence globale de significativité. Par conséquent, nous rejetons l'hypothèse d'absence globale de significativité et nous en déduisons que le modèle 1 est globalement avéré au seuil de 0,1%.

## 4.2 Analyse des résidus

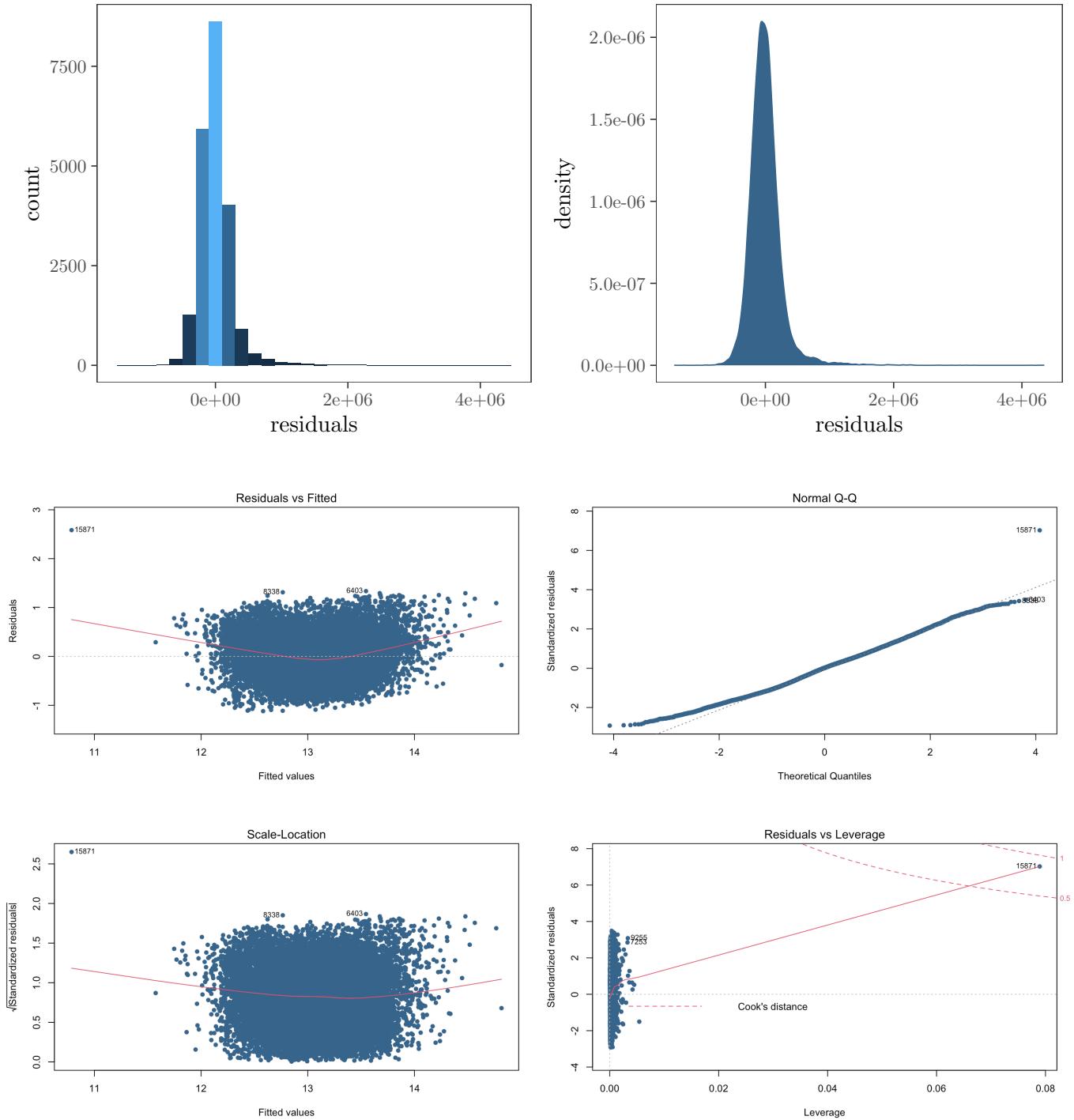


FIGURE 5 – Graphiques de l'analyse des résidus

TABLE 13 – Tests de l’analyse des résidus :

	Harrison_McCabe	Breusch_Pagan	Durbin_Watson
p-value	0.971	0.000	0.055
statistic	0.509	27.552	1.978

Pour  $DW$  nous avons 5,5% de chance de se tromper en rejetant l’hypothèse d’absence d’autocorrélation. Par conséquent, nous conservons l’hypothèse d’absence d’autocorrélation et nous en déduisons que les résidus ne sont pas autocorrélés au seuil de 0,1%.

Nous observons que les résidus sont globalement répartis le long de la droite. Nous en déduisons que les résidus suivent une loi normale.

Pour  $BP$ , nous avons moins de 0,1% de chance de se tromper en rejetant l’hypothèse d’homoscédasticité. Par conséquent, nous rejetons l’hypothèse d’homoscédasticité et nous en déduisons que les résidus sont hétéroscédastiques au seuil de 0,1%.

Nous observons l’existence de résidus qui ne sont globalement pas linéaires et qui dépassent la distance de Cook. Nous en déduisons que le logement 15871 et dans une moindre mesure les logements 9255, 8338, 7253 et 6403 ont un résidu aberrant.

```
data.analyse1 <- data[ - c( 15871,
                           9255,
                           8338,
                           7253,
                           6403 ), ]
```

Les hypothèses nécessaires des résultats et des résidus sont avérées au seuil de 0,1%. Notre constante, notre coefficient, notre significativité, notre linéarité des résidus, notre absence d’autocorrélation des résidus et notre appartenance à une loi normale des résidus sont avérées au seuil de 0,1%. L’hypothèse d’hétéroscédasticité ne pose pas de problème car le prix des logements et la variance du prix des logements ne sont pas homogènes.

En conclusion, de manière biaisé par la distance de Cook, nous avons 0,1% de chance de se tromper en affirmant que plus la surface intérieure en pieds carrés est élevée et plus le prix des logements est élevé selon la relation  $M_5$ .

### 4.3 Analyse du modèle 5 moins les résidus aberrants pour la distance de Cook

TABLE 14 – Régression linéaire du modèle 5 moins les résidus aberrants :

<i>Dependent variable:</i>	
	log(price)
log(sqft_living)	0.835*** (0.006)
Constant	6.740*** (0.047)
Observations	21,608
R <sup>2</sup>	0.455
Adjusted R <sup>2</sup>	0.455
Residual Std. Error	0.388 (df = 21606)
F Statistic	18,022.850*** (df = 1; 21606)

*Note:* \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

TABLE 15 – Performances de la régression linéaire du modèle 5 moins les résidus aberrants :

Name	Model	AIC	BIC	R2	R2_adjusted	RMSE	Sigma	Performance_Score
model5	lm	20444.54	20468.48	0.455	0.455	0.388	0.388	NaN

#### 4.3.1 Analyse des résultats

$$\mathcal{M}_5 : \log(y_i) = \alpha + \beta_6 \log(x_{i,6}) + z_i \iff \mathcal{M}_5 : \log(y_i) = 6,74 + 0,835 \log(x_{i,6}) + z_i$$

Pour  $\alpha$ , nous avons moins de 0,1% de chance de se tromper en rejetant l'hypothèse nulle. Par conséquent, nous rejetons l'hypothèse nulle et nous en déduisons que la constante du modèle 5 est différente de 0 au seuil de 0,1%.

Pour  $\beta_6$ , nous avons moins de 0,1% de chance de se tromper en rejetant l'hypothèse nulle. Par conséquent, nous rejetons l'hypothèse nulle et nous en déduisons que le coefficient du modèle 5 de la variable explicative `sqft_living` est différent de 0 au seuil de 0,1%.

Nous avons  $R^2 = 0,455$  dans le modèle 5 donc 45,5% de la variation du prix des logements est expliquée par la surface intérieure de ces derniers.

Pour  $F$ , nous avons moins de 0,1% de chance de se tromper en rejetant l'hypothèse d'absence globale de significativité. Par conséquent, nous rejetons l'hypothèse d'absence globale de significativité et nous en déduisons que le modèle 1 est globalement avéré au seuil de 0,1%.

#### 4.3.2 Analyse des résidus

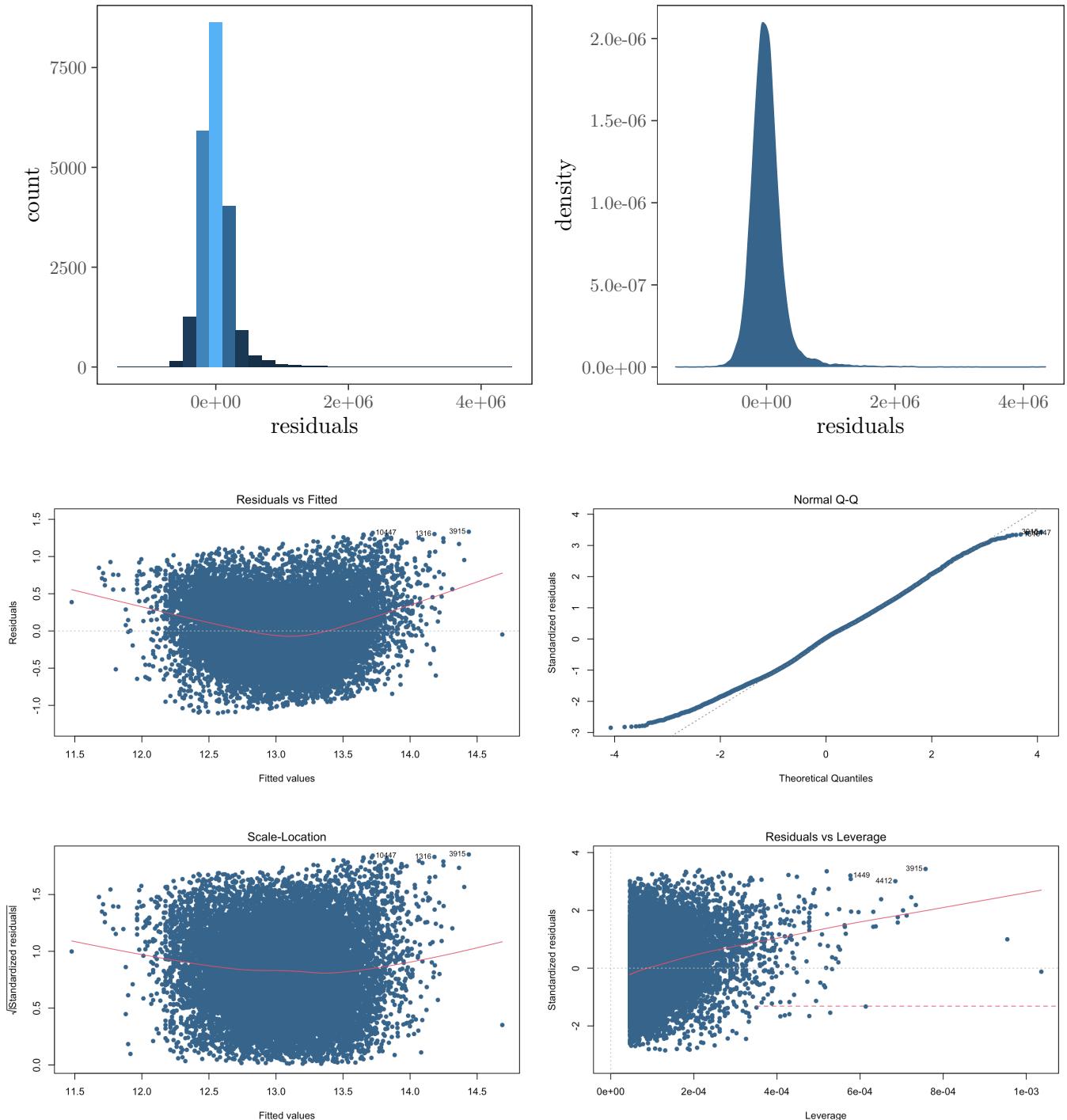


FIGURE 6 – Graphiques de l'analyse des résidus

TABLE 16 – Tests de l’analyse des résidus :

	Harrison_McCabe	Breusch_Pagan	Durbin_Watson
p-value	0.966	0.000	0.053
statistic	0.508	23.913	1.978

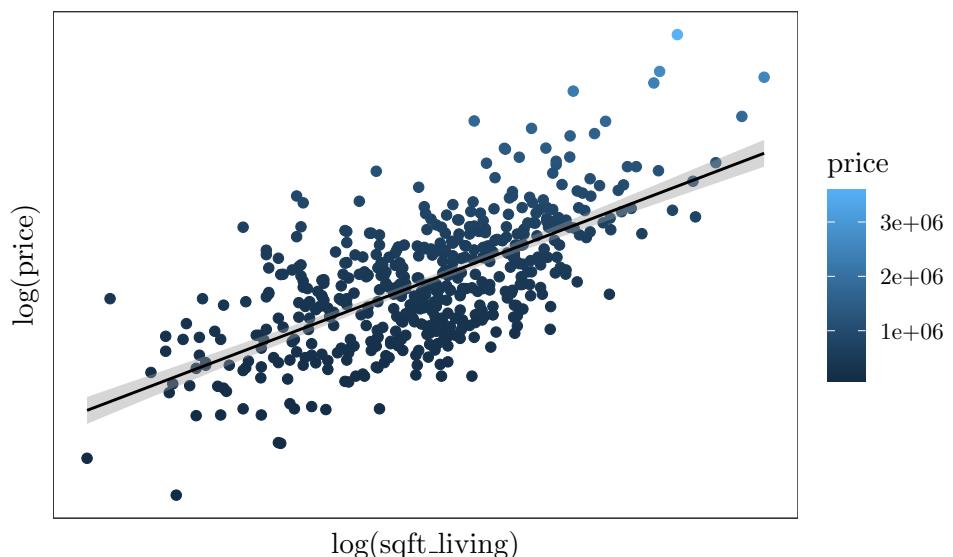
Pour  $DW$  nous avons 5,3% de chance de se tromper en rejetant l’hypothèse d’absence d’autocorrélation. Par conséquent, nous conservons l’hypothèse d’absence d’autocorrélation et nous en déduisons que les résidus ne sont pas autocorrélés au seuil de 0,1%.

Nous observons que les résidus sont globalement répartis le long de la droite. Nous en déduisons que les résidus suivent une loi normale.

Pour  $BP$ , nous avons moins de 0,1% de chance de se tromper en rejetant l’hypothèse d’homoscédasticité. Par conséquent, nous rejetons l’hypothèse d’homoscédasticité et nous en déduisons que les résidus sont hétéroscédastiques au seuil de 0,1%.

Nous observons l’existence de résidus qui ne sont globalement pas linéaires et qui tendent vers la distance de Cook. Nous en déduisons que, dans une moindre mesure, les logements 10447, 4412, 3915, 1449 et 1316 ont un résidu aberrant.

Nous avons 0,1% de chance de se tromper en affirmant que plus la surface intérieure en pieds carrés est élevée et plus le prix des logements est élevé selon la relation  $\mathcal{M}_5$ .



#### 4.4 $\mathcal{A}_1^H$ : l'analyse de l'hétéroscédasticité

Nous allons essayer de conserver l'hypothèse d'homoscédasticité des résidus en segmentant le prix des logements. Nous allons créer de nouvelles bases de données en fonction du minimum, du maximum et de la médiane :

```
data.analyse1.min_q1 <- data.analyse1[ data.analyse1$price >= 75000 &
                                         data.analyse1$price <= 321950, ]# modèle 1,
data.analyse1.min_q2 <- data.analyse1[ data.analyse1$price >= 75000 &
                                         data.analyse1$price <= 450000, ]# modèle 2,
data.analyse1.min_q3 <- data.analyse1[ data.analyse1$price >= 75000 &
                                         data.analyse1$price <= 645000, ]# modèle 3,
data.analyse1.q1_q2 <- data.analyse1[ data.analyse1$price >= 321950 &
                                         data.analyse1$price <= 450000, ]# modèle 4,
data.analyse1.q2_q3 <- data.analyse1[ data.analyse1$price >= 450000 &
                                         data.analyse1$price <= 645000, ]# modèle 5,
data.analyse1.q1_q3 <- data.analyse1[ data.analyse1$price >= 321950 &
                                         data.analyse1$price <= 645000, ]# modèle 6,
```

Nous pensons que les logements les plus chers ont une influence sur l'homogénéité du prix de ces derniers. En effet, le prix d'un logement varie grandement lorsque ce dernier est très élevé. Quelle est la différence entre un logement coûtant 1 000 000 dollars et un logement coûtant 1 500 000 dollars ? D'après les interviews *Asking Price* de la chaîne YouTube **Architectural Digest**, nous nous trompons communément sur le prix d'un logement en fonction de sa surface intérieure, de sa surface extérieure, etc. En effet, nous omettons le *standing*. Par conséquent, nous allons segmenter la base de données `BDD_data.csv` et comparer les résultats et les résidus des nouvelles bases de données.

$$\mathcal{A}_1^H = \{\mathcal{M}_1^H, \mathcal{M}_2^H, \mathcal{M}_3^H, \mathcal{M}_4^H, \mathcal{M}_5^H, \mathcal{M}_6^H\}$$

TABLE 17 – Régressions linéaires supplémentaires de l’analyse 1 :

Dependent variable:			
	log(price)		
	(1)	(2)	(3)
log(sqft_living)	0.276*** (0.008)	0.301*** (0.007)	0.447*** (0.007)
Constant	10.416*** (0.059)	10.427*** (0.052)	9.508*** (0.050)
Observations	5,404	10,864	16,239
R <sup>2</sup>	0.174	0.144	0.212
Adjusted R <sup>2</sup>	0.174	0.144	0.212
Residual Std. Error	0.200 (df = 5402)	0.259 (df = 10862)	0.313 (df = 16237)
F Statistic	1,140.195*** (df = 1; 5402)	1,828.210*** (df = 1; 10862)	4,376.828*** (df = 1; 16237)

Note:

\*p<0.1; \*\*p<0.05; \*\*\*p<0.01

TABLE 18 – Régressions linéaires supplémentaires de l’analyse 1 :

Dependent variable:			
	log(price)		
	(1)	(2)	(3)
log(sqft_living)	0.019*** (0.004)	0.067*** (0.004)	0.153*** (0.005)
Constant	12.716*** (0.029)	12.679*** (0.032)	11.873*** (0.039)
Observations	5,461	5,547	10,836
R <sup>2</sup>	0.004	0.044	0.073
Adjusted R <sup>2</sup>	0.004	0.044	0.073
Residual Std. Error	0.100 (df = 5459)	0.101 (df = 5545)	0.188 (df = 10834)
F Statistic	24.103*** (df = 1; 5459)	254.578*** (df = 1; 5545)	852.732*** (df = 1; 10834)

Note:

\*p<0.1; \*\*p<0.05; \*\*\*p<0.01

TABLE 19 – Comparaison des performances des régressions linéaires supplémentaires de l’analyse 1

Name	Model	AIC	BIC	R2	R2_adjusted	RMSE	Sigma	Performance_Score
model5	lm	-9655.489	-9635.626	0.044	0.044	0.101	0.101	0.727
model4	lm	-9668.215	-9648.399	0.004	0.004	0.100	0.100	0.667
model1	lm	-2047.937	-2028.152	0.174	0.174	0.200	0.200	0.642
model6	lm	-5422.781	-5400.909	0.073	0.073	0.188	0.188	0.560
model2	lm	1499.602	1521.482	0.144	0.144	0.259	0.259	0.435
model3	lm	8385.804	8408.889	0.212	0.212	0.313	0.313	0.333

TABLE 20 – Tests de l'analyse des résidus du modèle 1 :

	Harrison_McCabe	Breusch_Pagan	Durbin_Watson
p-value	0.017	0.000	0.263
statistic	0.479	587.741	1.983

TABLE 21 – Tests de l'analyse des résidus du modèle 2 :

	Harrison_McCabe	Breusch_Pagan	Durbin_Watson
p-value	0.855	0.000	0.319
statistic	0.507	1136.047	1.991

TABLE 22 – Tests de l'analyse des résidus du modèle 3 :

	Harrison_McCabe	Breusch_Pagan	Durbin_Watson
p-value	0.935	0.000	0.325
statistic	0.509	1010.537	1.993

TABLE 23 – Tests de l'analyse des résidus du modèle 4 :

	Harrison_McCabe	Breusch_Pagan	Durbin_Watson
p-value	0.745	0.312	0.365
statistic	0.507	1.023	1.991

TABLE 24 – Tests de l'analyse des résidus du modèle 5 :

	Harrison_McCabe	Breusch_Pagan	Durbin_Watson
p-value	0.536	0.001	0.895
statistic	0.501	11.256	2.034

TABLE 25 – Tests de l'analyse des résidus du modèle 6 :

	Harrison_McCabe	Breusch_Pagan	Durbin_Watson
p-value	0.521	0.00	0.556
statistic	0.500	70.24	2.003

En conclusion, nous conservons l'hypothèse d'homoscédasticité et nous en déduisons que les résidus sont homoscédastiques au seuil de 0,1% uniquement dans le modèle 4 c'est à dire pour les logements coûtant entre 321 950 dollars et 450 000 dollars. Est-ce que le prix d'un logement varie grandement même lorsque ce dernier n'est pas très élevé ? *A priori*, nous supposons que oui concernant la ville de Seattle dans l'État de Washington aux États-Unis.

5  $\mathcal{A}_2$  : l'analyse du prix d'un logement en fonction de son nombre de chambres, de son nombre de salles de bain et de son nombre d'étages

$$\mathcal{A}_2 = \{\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4\}$$

TABLE 26 – Régressions linéaires de l'analyse 2 :

	<i>Dependent variable:</i>	
	log(price)	price
	(1)	(2)
bedrooms	0.049*** (0.004)	20,024.380*** (2,680.838)
bathrooms	0.327*** (0.005)	238,461.400*** (3,681.888)
floors	0.055*** (0.006)	-1,737.472 (4,569.452)
Constant	12.109*** (0.013)	-29,102.610*** (9,209.213)
Observations	21,613	21,613
R <sup>2</sup>	0.311	0.278
Adjusted R <sup>2</sup>	0.310	0.278
Residual Std. Error (df = 21609)	0.437	312,040.000
F Statistic (df = 3; 21609)	3,244.010***	2,769.094***

Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

TABLE 27 – Régressions linéaires de l'analyse 2 :

	<i>Dependent variable:</i>	
	log(price)	price
	(1)	(2)
log(sqft_living)	0.839*** (0.010)	
sqft_living		309.393*** (3.087)
bedrooms	-0.070*** (0.004)	-57,847.960*** (2,347.323)
bathrooms	0.053*** (0.006)	7,853.522** (3,814.223)
floors	0.044*** (0.006)	200.497 (3,775.505)
Constant	6.772*** (0.067)	74,669.670*** (7,679.122)
Observations	21,613	21,613
R <sup>2</sup>	0.471	0.507
Adjusted R <sup>2</sup>	0.470	0.507
Residual Std. Error (df = 21608)	0.383	257,819.300
F Statistic (df = 4; 21608)	4,801.148***	5,553.623***

Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

## 5.1 Analyse des résultats

TABLE 28 – Aperçu de la base de données de l'analyse 2 :

	price	bedrooms	bathrooms	sqft_living	floors
19054	879950	4	2.25	3500	1.0
5372	350000	2	1.00	1620	1.0
20600	376000	3	2.00	1340	3.0
12819	1313000	6	3.00	2980	1.5
14746	355000	4	2.25	2200	2.0
9490	238000	5	2.25	2240	2.0
2560	890000	4	2.75	2610	1.0
14990	840000	4	3.25	3160	2.0
642	305000	4	2.50	2250	1.0
8372	435000	3	2.25	1890	1.0

TABLE 29 – Comparaison des performances des régressions linéaires de l'analyse 2

Name	Model	AIC	BIC	R2	R2 adjusted	RMSE	Sigma	Performance Score
model3	lm	19887.24	19935.13	0.471	0.470	0.383	0.383	0.947
model1	lm	25593.43	25633.33	0.311	0.310	0.437	0.437	0.711
model4	lm	599938.62	599986.51	0.507	0.507	257789.524	257819.348	0.396
model2	lm	608188.26	608228.17	0.278	0.278	312011.094	312039.970	0.000

Nous allons analyser le modèle 1.

$$\mathcal{M}_1 : \log(y_i) = \alpha + \beta_4 x_{i,4} + \beta_5 x_{i,5} + \beta_8 x_{i,8} + z_i \iff \mathcal{M}_1 : \log(y_i) = \alpha + 0,049x_{i,4} + 0,327x_{i,5} + 0,055x_{i,8} + z_i$$

Pour  $\alpha$ , nous avons moins de 0,1% de chance de se tromper en rejetant l'hypothèse nulle. Par conséquent, nous rejetons l'hypothèse nulle et nous en déduisons que la constante du modèle 1 est différente de 0 au seuil de 0,1%.

Pour  $\beta_4$ , nous avons moins de 0,1% de chance de se tromper en rejetant l'hypothèse nulle. Par conséquent, nous rejetons l'hypothèse nulle et nous en déduisons que le coefficient du modèle 1 de la variable explicative **bedrooms** est différent de 0 au seuil de 0,1%.

Pour  $\beta_5$ , nous avons moins de 0,1% de chance de se tromper en rejetant l'hypothèse nulle. Par conséquent, nous rejetons l'hypothèse nulle et nous en déduisons que le coefficient du modèle 1 de la variable explicative **bathrooms** est différent de 0 au seuil de 0,1%.

Pour  $\beta_8$ , nous avons moins de 0,1% de chance de se tromper en rejetant l'hypothèse nulle. Par conséquent, nous rejetons l'hypothèse nulle et nous en déduisons que le coefficient du modèle 1 de la variable explicative **floors** est différent de 0 au seuil de 0,1%.

Nous avons  $R^2 = 0,311$  dans le modèle 1 donc 31,1% de la variation du prix des logements est expliquée par le nombre de chambres, le nombre de salles de bain et le nombre d'étages de ces derniers.

Pour  $F$ , nous avons moins de 0,1% de chance de se tromper en rejetant l'hypothèse d'absence globale de significativité. Par conséquent, nous rejetons l'hypothèse d'absence globale de significativité et nous en déduisons que le modèle 1 est globalement avéré au seuil de 0,1%.

## 5.2 Analyse des résidus

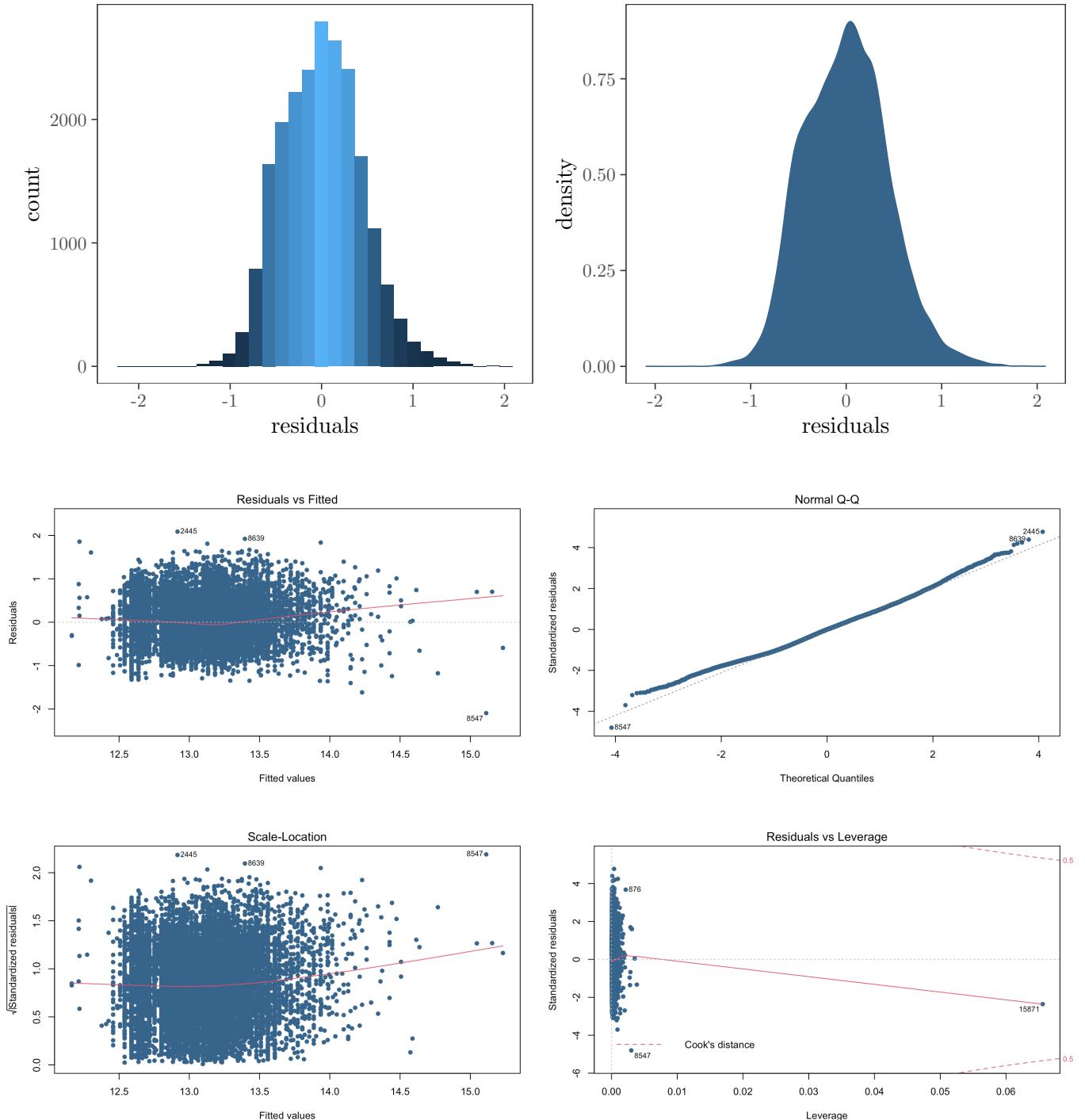


FIGURE 7 – Graphiques de l'analyse des résidus

TABLE 30 – Tests de l’analyse des résidus :

	Harrison_McCabe	Breusch_Pagan	Durbin_Watson
p-value	0.860	0.000	0.001
statistic	0.505	147.128	1.958

Pour *DW*, nous avons moins de 0,1% de chance de se tomper en rejetant l’hypothèse d’absence d’autocorrélation. Par conséquent, nous rejetons l’hypothèse d’absence d’autocorrélation et nous en déduisons que les résidus sont autocorrélés au seuil de 0,1%.

Nous observons que les résidus sont globalement répartis le long de la droite. Nous en déduisons que les résidus suivent une loi normale.

Pour *BP*, nous avons moins de 0,1% de chance de se tomper en rejetant l’hypothèse d’homoscédasticité. Par conséquent, nous rejetons l’hypothèse d’homoscédasticité et nous en déduisons que les résidus sont hétéroscédastiques au seuil de 0,1%.

Nous observons l’existence de résidus qui ne sont globalement pas linéaires et qui tendent vers la distance de Cook. Nous en déduisons que, dans une moindre mesure, les logements 15871, 8639, 8547, 2445 et 876 ont un résidu aberrant.

```
data.analyse2 <- data[ - c( 15871,
                           8639,
                           8547,
                           2445,
                           876 ), ]
```

Les hypothèses nécessaires des résultats sont avérés au seuil de 0,1%. Néanmoins les hypothèses nécessaires des résidus ne sont pas avérés au seuil de 0,1%. L’hypothèse d’hétéroscédasticité ne pose pas de problème car le prix des logements et la variance du prix des logements ne sont pas homogènes mais l’hypothèse d’autocorrélation pose un problème car les résidus ne sont pas linéaires et aléatoires.

### 5.3 $\mathcal{A}_2^H$ : l'analyse de l'autocorrélation

Nous allons essayer de conserver l'hypothèse d'autocorrélation des résidus en segmentant le prix des logements. Nous allons créer de nouvelles bases de données en fonction du minimum, du maximum et de la médiane :

```
data.analyse2.min_q1 <- data.analyse2[ data.analyse2$price >= 75000 &
                                         data.analyse2$price <= 321950, ]# modèle 1,
data.analyse2.min_q2 <- data.analyse2[ data.analyse2$price >= 75000 &
                                         data.analyse2$price <= 450000, ]# modèle 2,
data.analyse2.min_q3 <- data.analyse2[ data.analyse2$price >= 75000 &
                                         data.analyse2$price <= 645000, ]# modèle 3,
data.analyse2.q1_q2 <- data.analyse2[ data.analyse2$price >= 321950 &
                                         data.analyse2$price <= 450000, ]# modèle 4,
data.analyse2.q2_q3 <- data.analyse2[ data.analyse2$price >= 450000 &
                                         data.analyse2$price <= 645000, ]# modèle 5,
data.analyse2.q1_q3 <- data.analyse2[ data.analyse2$price >= 321950 &
                                         data.analyse2$price <= 645000, ]# modèle 6,
```

$$\mathcal{A}_1^A = \{\mathcal{M}_1^A, \mathcal{M}_2^A, \mathcal{M}_3^A, \mathcal{M}_4^A, \mathcal{M}_5^A, \mathcal{M}_6^A\}$$

TABLE 31 – Régressions linéaires supplémentaire de l'analyse 2 :

	<i>Dependent variable:</i>		
	log(price)		
	(1)	(2)	(3)
bedrooms	0.023*** (0.004)	0.015*** (0.003)	0.025*** (0.003)
bathrooms	0.106*** (0.006)	0.116*** (0.005)	0.161*** (0.005)
floors	0.044*** (0.007)	0.046*** (0.006)	0.048*** (0.006)
Constant	12.110*** (0.012)	12.318*** (0.011)	12.362*** (0.011)
Observations	5,404	10,863	16,238
R <sup>2</sup>	0.156	0.115	0.142
Adjusted R <sup>2</sup>	0.156	0.115	0.142
Residual Std. Error	0.202 (df = 5400)	0.264 (df = 10859)	0.327 (df = 16234)
F Statistic	333.772*** (df = 3; 5400)	472.413*** (df = 3; 10859)	893.796*** (df = 3; 16234)

Note:

\*p<0.1; \*\*p<0.05; \*\*\*p<0.01

TABLE 32 – Régressions linéaires supplémentaires de l'analyse 2 :

	<i>Dependent variable:</i>		
	log(price)		
	(1)	(2)	(3)
bedrooms	-0.004** (0.002)	0.009*** (0.002)	0.008*** (0.002)
bathrooms	0.011*** (0.003)	0.011*** (0.003)	0.049*** (0.004)
floors	-0.006** (0.003)	0.008*** (0.003)	-0.001 (0.004)
Constant	12.857*** (0.006)	13.122*** (0.007)	12.897*** (0.009)
Observations	5,460	5,546	10,835
R <sup>2</sup>	0.003	0.019	0.033
Adjusted R <sup>2</sup>	0.002	0.019	0.033
Residual Std. Error	0.100 (df = 5456)	0.103 (df = 5542)	0.192 (df = 10831)
F Statistic	5.536*** (df = 3; 5456)	36.286*** (df = 3; 5542)	122.607*** (df = 3; 10831)

Note:

\*p<0.1; \*\*p<0.05; \*\*\*p<0.01

TABLE 33 – Comparaison des performances des régressions linéaires supplémentaires de l'analyse 2

Name	Model	AIC	BIC	R2	R2_adjusted	RMSE	Sigma	Performance_Score
model1	lm	-1928.296	-1895.322	0.156	0.156	0.202	0.202	0.717
model5	lm	-9510.411	-9477.307	0.019	0.019	0.103	0.103	0.695
model4	lm	-9656.521	-9623.495	0.003	0.002	0.100	0.100	0.667
model6	lm	-4958.215	-4921.762	0.033	0.033	0.192	0.192	0.515
model2	lm	1859.896	1896.362	0.115	0.115	0.263	0.264	0.474
model3	lm	9783.224	9821.699	0.142	0.142	0.327	0.327	0.302

TABLE 34 – Tests de l'analyse des résidus du modèle 1 :

	Harrison_McCabe	Breusch_Pagan	Durbin_Watson
p-value	0.096	0.000	0.410
statistic	0.487	407.933	1.994

TABLE 35 – Tests de l'analyse des résidus du modèle 2 :

	Harrison_McCabe	Breusch_Pagan	Durbin_Watson
p-value	0.917	0.000	0.418
statistic	0.510	849.865	1.996

TABLE 36 – Tests de l'analyse des résidus du modèle 3 :

	Harrison_McCabe	Breusch_Pagan	Durbin_Watson
p-value	0.980	0.000	0.183
statistic	0.512	802.932	1.986

TABLE 37 – Tests de l'analyse des résidus du modèle 4 :

	Harrison_McCabe	Breusch_Pagan	Durbin_Watson
p-value	0.706	0.553	0.336
statistic	0.505	2.093	1.989

TABLE 38 – Tests de l'analyse des résidus du modèle 5 :

	Harrison_McCabe	Breusch_Pagan	Durbin_Watson
p-value	0.474	0.062	0.872
statistic	0.499	7.343	2.031

TABLE 39 – Tests de l'analyse des résidus du modèle 6 :

	Harrison_McCabe	Breusch_Pagan	Durbin_Watson
p-value	0.385	0.000	0.480
statistic	0.498	51.599	1.999

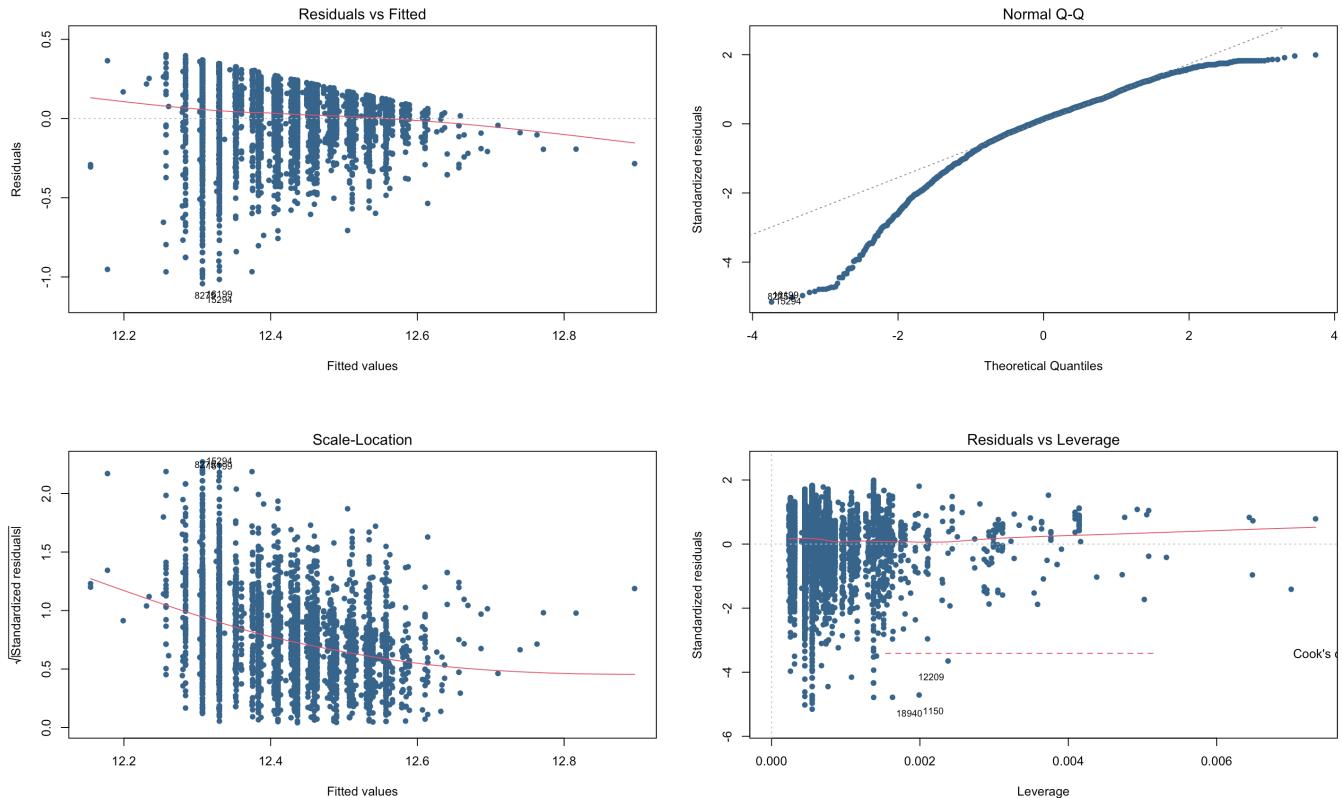


FIGURE 8 – Graphiques de l'analyse des résidus

$$\mathcal{M}_1^A : \log(y_i) = \alpha + \beta_4 x_{i,4} + \beta_5 x_{i,5} + \beta_8 x_{i,8} + z_i \iff \mathcal{M}_1^A : \log(y_i) = 12,11 + 0,023 x_{i,4} + 0,106 x_{i,5} + 0,044 x_{i,8} + z_i$$

En conclusion, nous conservons les hypothèses d'absence d'autocorrélation et nous en déduisons que les résidus ne sont pas autocorrélés au seuil de 0,1% dans les modèles 1, 2, 3, 4, 5 et 6. De manière biaisé par les graphiques de l'analyse des résidus qui ne sont pas bons, nous avons 0,1% de chance de se tromper en affirmant que plus le nombre de chambres, le nombre de salles de bain et le nombre d'étages sont élevés et plus le prix des logements est élevé selon la relation  $\mathcal{M}_1^A$ .

## 6 $\mathcal{A}_3$ : l'analyse du prix d'un logement à partir de 645 000 dollars en fonction de sa vue sur la mer ou non, de son point de vue et de son *design*

Nous allons essayer d'analyser le *standing* dont nous avons parlé dans le 4.4. Par conséquent, nous allons segmenter la base de données `BDD_data.csv` en ne prenant en compte que les logements coûtant plus de 645 000 dollars.

```
data.analyse3 <- data[ data$price >= 645000, ]
```

$$\mathcal{A}_3 = \{\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4\}$$

TABLE 40 – Régressions linéaires de l'analyse 3 :

	Dependent variable:	
	log(price)	price
	(1)	(2)
waterfront2	0.332*** (0.027)	600,302.300*** (37,981.760)
view2	0.130*** (0.019)	166,096.600*** (26,035.900)
view3	0.096*** (0.012)	104,150.500*** (16,487.350)
view4	0.172*** (0.014)	190,977.600*** (20,029.270)
view5	0.238*** (0.019)	288,291.500*** (27,054.850)
grade5	0.182 (0.134)	230,233.800 (186,647.300)
grade6	0.220* (0.127)	306,892.500* (176,938.000)
grade7	0.278** (0.127)	356,616.400** (176,638.200)
grade8	0.382*** (0.127)	464,067.200*** (176,627.800)
grade9	0.540*** (0.127)	651,744.500*** (176,730.900)
grade10	0.789*** (0.127)	996,335.500*** (177,236.000)
grade11	1.094*** (0.129)	1,591,935.000*** (180,065.100)
grade12	1.644*** (0.145)	3,174,180.000*** (201,539.200)
Constant	13.280*** (0.127)	415,491.000** (176,509.000)
Observations	5,413	5,413
R <sup>2</sup>	0.444	0.454
Adjusted R <sup>2</sup>	0.443	0.452
Residual Std. Error (df = 5399)	0.253	352,169.000
F Statistic (df = 13; 5399)	331.901***	344.646***

Note:

\*p<0.1; \*\*p<0.05; \*\*\*p<0.01

TABLE 41 – Régressions linéaires de l'analyse 3 :

	<i>Dependent variable:</i>	
	log(price)	price
	(1)	(2)
log(sqft_living)	0.310*** (0.014)	
sqft_living		177.778*** (6.344)
waterfront2	0.329*** (0.026)	589,476.800*** (35,493.090)
view2	0.119*** (0.018)	145,610.000*** (24,339.490)
view3	0.084*** (0.011)	82,137.050*** (15,426.160)
view4	0.155*** (0.014)	157,060.900*** (18,754.880)
view5	0.217*** (0.019)	244,800.600*** (25,328.230)
grade5	0.106 (0.129)	158,734.500 (174,426.000)
grade6	0.032 (0.122)	114,381.100 (165,477.400)
grade7	0.030 (0.122)	92,888.610 (165,322.700)
grade8	0.073 (0.122)	106,255.900 (165,538.000)
grade9	0.180 (0.123)	199,861.400 (165,926.500)
grade10	0.365*** (0.123)	398,163.300** (166,982.900)
grade11	0.609*** (0.126)	810,832.700*** (170,549.600)
grade12	1.064*** (0.141)	2,036,456.000*** (192,648.700)
Constant	11.114*** (0.156)	243,224.000 (165,048.300)
Observations	5,413	5,413
R <sup>2</sup>	0.490	0.523
Adjusted R <sup>2</sup>	0.489	0.522
Residual Std. Error (df = 5398)	0.242	329,074.400
F Statistic (df = 14; 5398)	370.816***	422.624***

Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

## 6.1 Analyse des résultats

TABLE 42 – Comparaison des performances des régressions linéaires de l'analyse 3

Name	Model	AIC	BIC	R2	R2_adjusted	RMSE	Sigma	Performance_Score
model3	lm	36.558	142.103	0.490	0.489	0.242	0.242	0.862
model1	lm	502.793	601.741	0.444	0.443	0.253	0.253	0.666
model4	lm	152912.336	153017.881	0.523	0.522	328618.144	329074.410	0.357
model2	lm	153645.636	153744.584	0.454	0.452	351713.255	352168.968	0.039

$$\mathcal{M}_1 : \log(y_i) = \alpha + \beta_9 x_{i,9} + \beta_{10} x_{i,10} + \beta_{12} x_{i,12} + z_i$$

Nous allons analyser le modèle 1.

Pour  $\alpha$ , nous avons moins de 0,1% de chance de se tromper en rejetant l'hypothèse nulle. Par conséquent, nous rejetons l'hypothèse nulle et nous en déduisons que la constante du modèle 1 est différente de 0 au seuil de 0,1%.

Pour  $\beta_9^1$ , nous avons moins de 0,1% de chance de se tromper en rejetant l'hypothèse nulle. Par conséquent, nous rejetons l'hypothèse nulle et nous en déduisons que le coefficient du modèle 1 de la variable explicative `waterfront1` est différent de 0 au seuil de 0,1%.

Pour  $\beta_{10}^{2,3,4,5}$ , nous avons moins de 0,1% de chance de se tromper en rejetant l'hypothèse nulle. Par conséquent, nous rejetons l'hypothèse nulle et nous en déduisons que les coefficients du modèle 1 des variables explicatives `view2`, `view3`, `view4` et `view5` sont différents de 0 au seuil de 0,1%.

Pour  $\beta_{12}^{6,7,8,9}$ , nous avons entre 0,3% et 17,6% de chance de se tromper en rejetant l'hypothèse nulle. Par conséquent, nous conservons l'hypothèse nulle et nous en déduisons que les coefficients du modèle 1 des variables explicatives `grade6`, `grade7`, `grade8` et `grade9` sont égales à 0 au seuil de 0,1%.

Pour  $\beta_{12}^{10,11,12,13}$ , nous avons moins de 0,1% de chance de se tromper en rejetant l'hypothèse nulle. Par conséquent, nous rejetons l'hypothèse nulle et nous en déduisons que les coefficients du modèle 1 des variables explicatives `grade10`, `grade11`, `grade12` et `grade13` sont différents de 0 au seuil de 0,1%.

Nous avons  $R^2 = 0.444$  dans le modèle 1 donc 44,4% de la variation du prix des logements est expliquée par le point de vue, la vue sur la mer et le *design* de ces derniers.

Pour  $F$ , nous avons moins de 0,1% de chance de se tromper en rejetant l'hypothèse d'absence globale de significativité. Par conséquent, nous rejetons l'hypothèse d'absence globale de significativité et nous en déduisons que le modèle 1 est globalement avéré au seuil de 0,1%.

## 6.2 Analyse des résidus

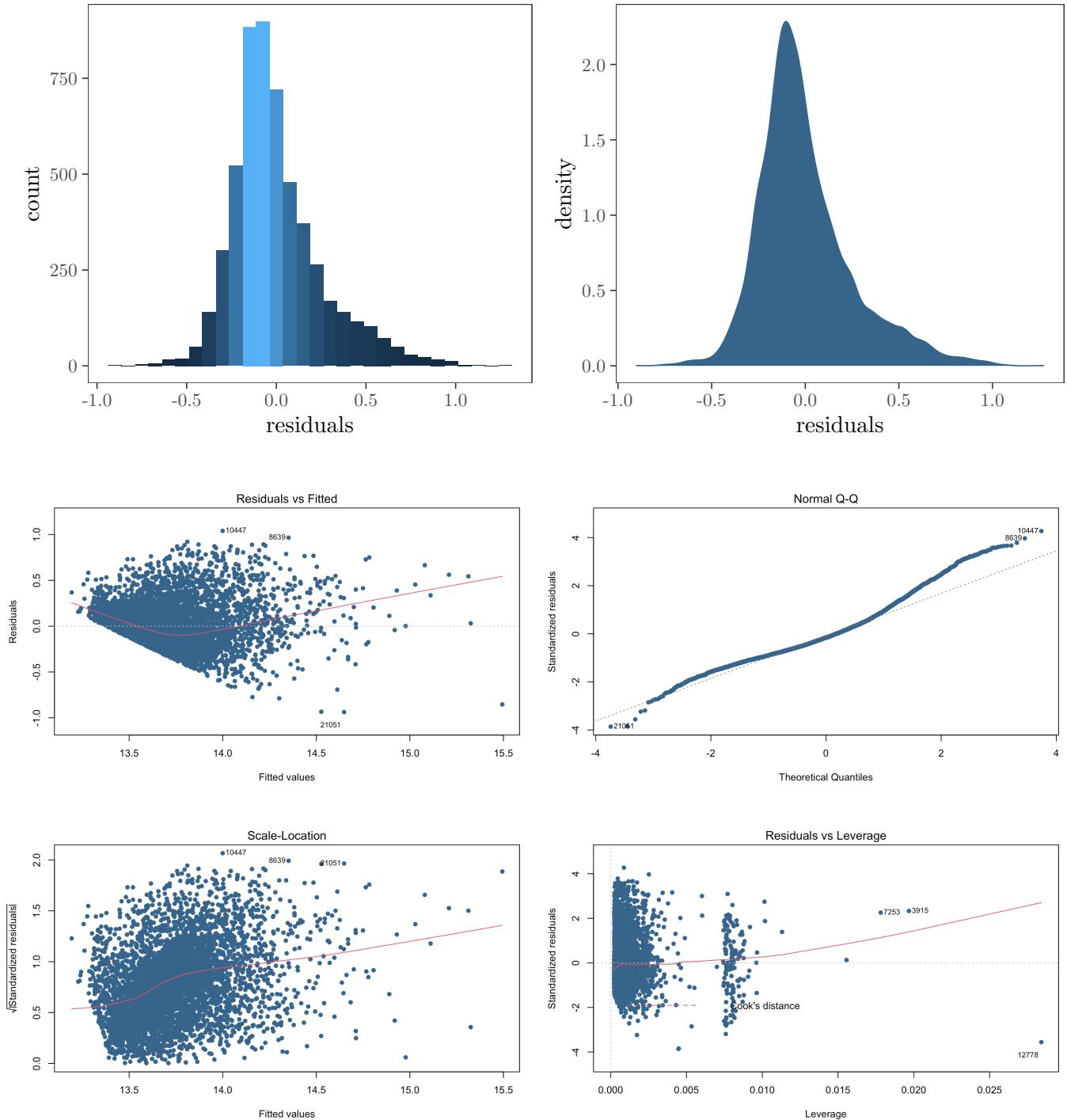


FIGURE 9 – Graphiques de l'analyse des résidus

TABLE 43 – Tests de l’analyse des résidus :

	Harrison_McCabe	Breusch_Pagan	Durbin_Watson
p-value	0.079	0.000	0.084
statistic	0.486	462.261	1.963

Pour  $DW$ , nous avons 8,37% de chance de se tomper en rejetant l’hypothèse d’absence d’autocorrélation. Par conséquent, nous conservons l’hypothèse d’absence d’autocorrélation et nous en déduisons que les résidus ne sont pas autocorrélés au seuil de 0,1%.

Nous observons que les résidus ne sont globalement pas répartis le long de la droite. Néanmoins, nous en déduisons que les résidus suivent une loi normale.

Pour  $BP$ , nous avons moins de 0,1% de chance de se tomper en rejetant l’hypothèse d’homoscédasticité. Par conséquent, nous rejetons l’hypothèse d’homoscédasticité et nous en déduisons que les résidus sont hétéroscédastiques au seuil de 0,1%.

Nous observons l’existence de résidus qui ne sont globalement pas linéaires et qui tendent vers la distance de Cook. Nous en déduisons que, dans une moindre mesure, les logements 21051, 12778, 10447, 8639, 7253 et 3915 ont un résidu aberrant.

Les hypothèses nécessaires des résultats et des résidus sont avérés au seuil de 0,1%. Notre constante, notre coefficient, notre significativité, notre linéarité des résidus, notre absence d’autocorrélation des résidus et notre appartenance à une loi normale des résidus sont avérés au seuil de 0,1%. L’hypothèse d’hétéroscédasticité ne pose pas de problème car le prix des logements et la variance du prix des logements ne sont pas homogènes **notamment** pour les logements coûtant plus de 645 000 dollars.

En conclusion, nous avons 0,1% de chance de se tromper en affirmant que plus le point de vue est beau et plus le *design* est classe et plus le prix des logements les logements coûtant plus de 645 000 dollars est élevé selon la relation  $\mathcal{M}_1$ .

## 7 $\mathcal{A}_4$ : l'analyse du prix d'un logement en fonction de son état

TABLE 44 – Régressions linéaires de l'analyse 4 :

	<i>Dependent variable:</i>	
	log(price)	price
	(1)	(2)
condition2	0.047 (0.103)	-7,144.521 (72,395.060)
condition3	0.565*** (0.096)	207,580.900*** (66,874.610)
condition4	0.521*** (0.096)	186,768.700*** (66,979.450)
condition5	0.667*** (0.096)	277,986.400*** (67,389.750)
Constant	12.491*** (0.095)	334,431.700*** (66,803.230)
Observations	21,613	21,613
R <sup>2</sup>	0.014	0.007
Adjusted R <sup>2</sup>	0.014	0.007
Residual Std. Error (df = 21608)	0.523	365,896.400
F Statistic (df = 4; 21608)	75.880***	37.412***

Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

TABLE 45 – Régressions linéaires de l'analyse 4 :

	<i>Dependent variable:</i>	
	log(price)	price
	(1)	(2)
log(sqft_living)	0.837*** (0.006)	
sqft_living		282.642*** (1.938)
condition2	-0.089 (0.076)	-61,993.580 (51,394.280)
condition3	0.073 (0.071)	-56,136.330 (47,508.390)
condition4	0.101 (0.071)	-20,971.100 (47,569.710)
condition5	0.220*** (0.071)	49,919.100 (47,865.210)
Constant	6.635*** (0.083)	-9,261.488 (47,481.810)
Observations	21,613	21,613
R <sup>2</sup>	0.462	0.500
Adjusted R <sup>2</sup>	0.462	0.499
Residual Std. Error (df = 21607)	0.386	259,748.100
F Statistic (df = 5; 21607)	3,711.535***	4,313.423***

Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

## 7.1 Analyse des résultats

TABLE 46 – Comparaison des performances des régressions linéaires de l'analyse 4

Name	Model	AIC	BIC	R2	R2_adjusted	RMSE	Sigma	Performance_Score
model3	lm	20234.11	20289.98	0.462	0.462	0.386	0.386	0.975
model1	lm	33330.06	33377.94	0.014	0.014	0.523	0.523	0.664
model4	lm	600261.78	600317.65	0.500	0.499	259711.998	259748.055	0.438
model2	lm	615071.65	615119.54	0.007	0.007	365854.028	365896.354	0.000

Nous allons analyser le modèle 1.

$$\mathcal{M}_1 : \log(y_i) = \alpha + \beta_{11}x_{i,11} + z_i$$

Pour  $\alpha$ , nous avons moins de 0,1% de chance de se tromper en rejetant l'hypothèse nulle. Par conséquent, nous rejetons l'hypothèse nulle et nous en déduisons que la constante du modèle 1 est différente de 0 au seuil de 0,1%.

Pour  $\beta_{11}^2$ , nous avons 65% de chance de se tromper en rejetant l'hypothèse nulle. Par conséquent, nous conservons l'hypothèse nulle et nous en déduisons que le coefficient du modèle 1 de la variable explicative **condition2** est égale à 0 au seuil de 0,1%.

Pour  $\beta_{11}^{3,4,5}$ , nous avons moins de 0,1% de chance de se tromper en rejetant l'hypothèse nulle. Par conséquent, nous rejetons l'hypothèse nulle et nous en déduisons que les coefficients du modèle 1 des variables explicatives **condition3**, **condition4** et **condition5** sont différents de 0 au seuil de 0,1%.

Nous avons  $R^2 = 0.014$  dans le modèle 1 donc 1.4% de la variation du prix des logements est expliquée par l'état de ces derniers.

Pour  $F$ , nous avons moins de 0,1% de chance de se tromper en rejetant l'hypothèse d'absence globale de significativité. Par conséquent, nous rejetons l'hypothèse d'absence globale de significativité et nous en déduisons que le modèle 1 est globalement avéré au seuil de 0,1%.

## 7.2 Analyse des résidus

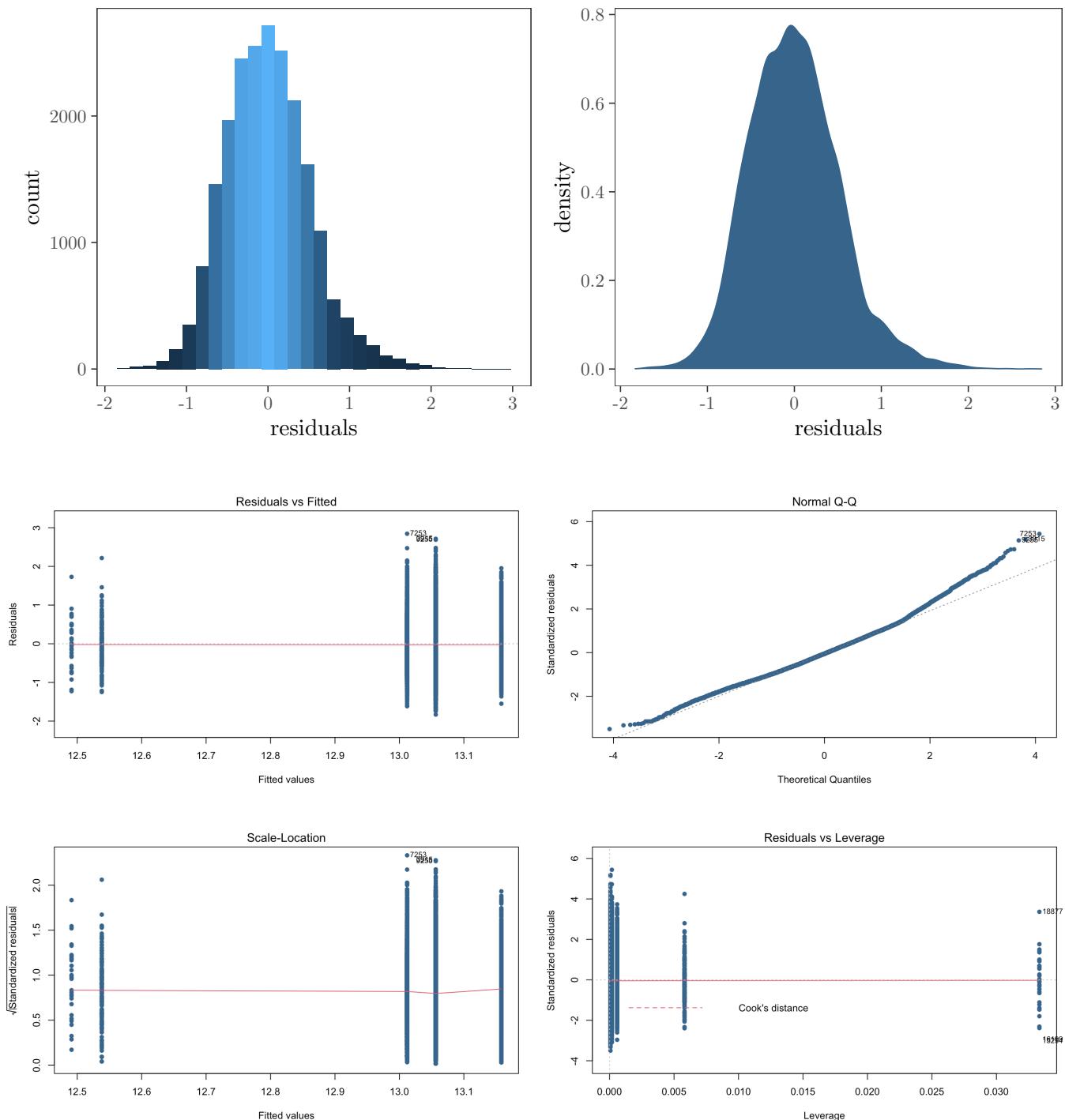


FIGURE 10 – Graphiques de l'analyse des résidus

TABLE 47 – Tests de l’analyse des résidus :

	Harrison_McCabe	Breusch_Pagan	Durbin_Watson
p-value	0.366	0.000	0.001
statistic	0.498	21.666	1.959

Pour  $DW$ , nous avons 0,1% de chance de se tomber en rejetant l’hypothèse d’absence d’autocorrélation. Par conséquent, nous rejetons l’hypothèse d’absence d’autocorrélation et nous en déduisons que les résidus sont autocorrélés au seuil de 0,1%.

Nous observons que les résidus ne sont globalement pas répartis le long de la droite. Néanmoins, nous en déduisons que les résidus suivent une loi normale.

Pour  $BP$ , nous avons moins de 0,1% de chance de se tomber en rejetant l’hypothèse d’homoscédasticité. Par conséquent, nous rejetons l’hypothèse d’homoscédasticité et nous en déduisons que les résidus sont hétéroscléastiques au seuil de 0,1%.

Nous observons l’existence de résidus qui ne sont globalement pas linéaires et qui tendent vers la distance de Cook. Nous en déduisons que, dans une moindre mesure, les logements 19294, 18877, 16199, 9255, 7253 et 3915 ont un résidu aberrant.

```
data.analyse4 <- data[ - c( 19294,
                           18877,
                           16199,
                           9255,
                           7253,
                           3915 ), ]
```

Les hypothèses nécessaires des résultats sont avérées au seuil de 0,1%. Néanmoins les hypothèses nécessaires des résidus ne sont pas avérées au seuil de 0,1%. L’hypothèse d’hétéroscléasticité ne pose pas de problème car le prix des logements et la variance du prix des logements ne sont pas homogènes mais l’hypothèse d’autocorrélation pose un problème car les résidus ne sont pas linéaires et aléatoires.

### 7.3 $\mathcal{A}_4^H$ : l'analyse de l'autocorrélation

Nous allons essayer de conserver l'hypothèse d'autocorrélation des résidus en segmentant le prix des logements. Nous allons créer de nouvelles bases de données en fonction du minimum, du maximum et de la médiane :

```
data.analyse4.min_q1 <- data.analyse4[ data.analyse4$price >= 75000 &
                                         data.analyse4$price <= 321950, ]# modèle 1,
data.analyse4.min_q2 <- data.analyse4[ data.analyse4$price >= 75000 &
                                         data.analyse4$price <= 450000, ]# modèle 2,
data.analyse4.min_q3 <- data.analyse4[ data.analyse4$price >= 75000 &
                                         data.analyse4$price <= 645000, ]# modèle 3,
data.analyse4.q1_q2 <- data.analyse4[ data.analyse4$price >= 321950 &
                                         data.analyse4$price <= 450000, ]# modèle 4,
data.analyse4.q2_q3 <- data.analyse4[ data.analyse4$price >= 450000 &
                                         data.analyse4$price <= 645000, ]# modèle 5,
data.analyse4.q1_q3 <- data.analyse4[ data.analyse4$price >= 321950 &
                                         data.analyse4$price <= 645000, ]# modèle 6,
```

$$\mathcal{A}_1^A = \{\mathcal{M}_1^A, \mathcal{M}_2^A, \mathcal{M}_3^A, \mathcal{M}_4^A, \mathcal{M}_5^A, \mathcal{M}_6^A\}$$

TABLE 48 – Régressions linéaires supplémentaires de l’analyse 4 :

	<i>Dependent variable:</i>		
	log(price)		
	(1)	(2)	(3)
condition2	0.119** (0.057)	0.023 (0.061)	0.033 (0.073)
condition3	0.320*** (0.053)	0.314*** (0.057)	0.397*** (0.067)
condition4	0.305*** (0.053)	0.274*** (0.057)	0.360*** (0.068)
condition5	0.311*** (0.054)	0.293*** (0.058)	0.407*** (0.068)
Constant	12.104*** (0.053)	12.339*** (0.057)	12.438*** (0.067)
Observations	5,402	10,862	16,238
R <sup>2</sup>	0.023	0.019	0.014
Adjusted R <sup>2</sup>	0.022	0.019	0.014
Residual Std. Error	0.217 (df = 5397)	0.277 (df = 10857)	0.350 (df = 16233)
F Statistic	31.480*** (df = 4; 5397)	53.388*** (df = 4; 10857)	59.000*** (df = 4; 16233)

Note:

\*p<0.1; \*\*p<0.05; \*\*\*p<0.01

TABLE 49 – Régressions linéaires supplémentaires de l’analyse 4 :

	<i>Dependent variable:</i>		
	log(price)		
	(1)	(2)	(3)
condition2	-0.075* (0.042)	-0.067 (0.064)	-0.036 (0.067)
condition3	-0.055 (0.038)	-0.037 (0.060)	0.015 (0.062)
condition4	-0.051 (0.038)	-0.036 (0.060)	0.025 (0.062)
condition5	-0.040 (0.038)	-0.015 (0.060)	0.053 (0.062)
Constant	12.911*** (0.038)	13.223*** (0.060)	13.004*** (0.062)
Observations	5,461	5,548	10,837
R <sup>2</sup>	0.002	0.003	0.003
Adjusted R <sup>2</sup>	0.001	0.003	0.002
Residual Std. Error	0.100 (df = 5456)	0.103 (df = 5543)	0.195 (df = 10832)
F Statistic	2.590** (df = 4; 5456)	4.740*** (df = 4; 5543)	7.714*** (df = 4; 10832)

Note:

\*p<0.1; \*\*p<0.05; \*\*\*p<0.01

TABLE 50 – Comparaison des performances des régressions linéaires supplémentaires de l'analyse 4

Name	Model	AIC	BIC	R2	R2_adjusted	RMSE	Sigma	Performance_Score
model1	lm	-1155.893	-1116.326	0.023	0.022	0.217	0.217	0.713
model5	lm	-9419.085	-9379.357	0.003	0.003	0.103	0.103	0.683
model4	lm	-9648.516	-9608.884	0.002	0.001	0.100	0.100	0.667
model2	lm	2962.730	3006.488	0.019	0.019	0.277	0.277	0.517
model6	lm	-4624.939	-4581.194	0.003	0.002	0.195	0.195	0.480
model3	lm	12017.053	12063.224	0.014	0.014	0.350	0.350	0.202

TABLE 51 – Tests de l'analyse des résidus du modèle 1 :

	Harrison_McCabe	Breusch_Pagan	Durbin_Watson
p-value	0.048	0.000	0.006
statistic	0.484	64.779	1.932

TABLE 52 – Tests de l'analyse des résidus du modèle 2 :

	Harrison_McCabe	Breusch_Pagan	Durbin_Watson
p-value	0.812	0.000	0.065
statistic	0.506	84.047	1.971

TABLE 53 – Tests de l'analyse des résidus du modèle 3 :

	Harrison_McCabe	Breusch_Pagan	Durbin_Watson
p-value	0.893	0.000	0.027
statistic	0.507	81.606	1.970

TABLE 54 – Tests de l'analyse des résidus du modèle 4 :

	Harrison_McCabe	Breusch_Pagan	Durbin_Watson
p-value	0.713	0.055	0.326
statistic	0.505	9.260	1.988

TABLE 55 – Tests de l'analyse des résidus du modèle 5 :

	Harrison_McCabe	Breusch_Pagan	Durbin_Watson
p-value	0.441	0.048	0.900
statistic	0.499	9.609	2.034

TABLE 56 – Tests de l'analyse des résidus du modèle 6 :

	Harrison_McCabe	Breusch_Pagan	Durbin_Watson
p-value	0.224	0.862	0.449
statistic	0.495	1.297	1.998

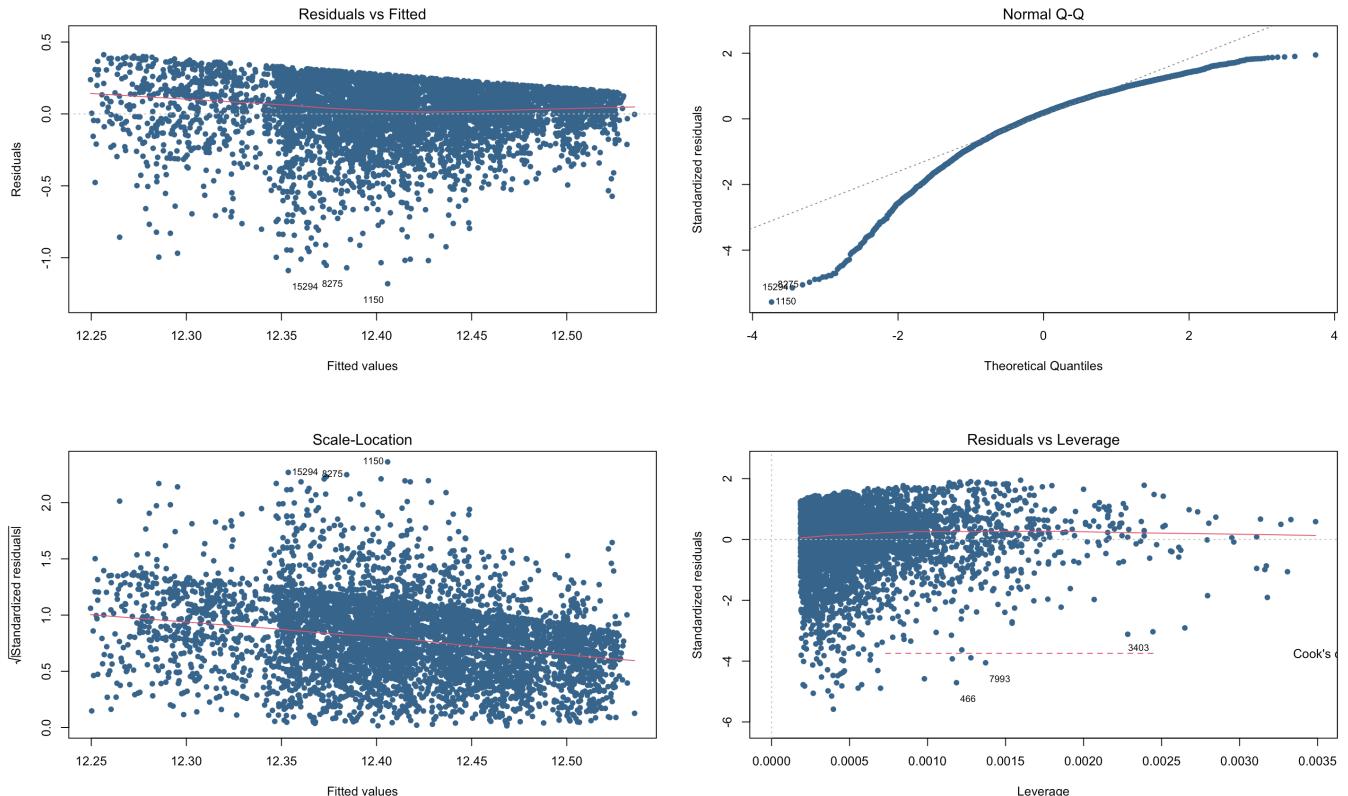


FIGURE 11 – Graphiques de l'analyse des résidus

En conclusion, nous conservons les hypothèses d'absence d'autocorrélation et nous en déduisons que les résidus ne sont pas autocorrélés au seuil de 0,1% dans les modèles 1, 2, 3, 4, 5 et 6. De manière biaisé par les graphiques de l'analyse des résidus qui ne sont pas bons, nous avons 1% de chance de se tromper en affirmant que plus les logements sont en bon état et plus le prix de ces derniers est élevé selon la relation  $\mathcal{M}_1^A$ .

## 8 $\mathcal{A}_5$ : l'analyse du prix d'un logement en fonction de son âge de construction et de son âge de rénovation

TABLE 57 – Régressions linéaires de l'analyse 5 :

	<i>Dependent variable:</i>	
	price	
	(1)	(2)
life_built	-674.369*** (84.893)	-429.877*** (131.956)
life_renovated	-65.659 (160.771)	510.952* (287.429)
life_built:life_renovated		-13.282** (5.488)
Constant	570,514.500*** (5,352.348)	559,881.500*** (6,924.413)
Observations	21,613	21,613
R <sup>2</sup>	0.003	0.003
Adjusted R <sup>2</sup>	0.003	0.003
Residual Std. Error	366,608.100 (df = 21610)	366,566.900 (df = 21609)
F Statistic	31.625*** (df = 2; 21610)	23.040*** (df = 3; 21609)

Note:

\*p<0.1; \*\*p<0.05; \*\*\*p<0.01

TABLE 58 – Régressions linéaires de l'analyse 5 :

	<i>Dependent variable:</i>	
	log(price)	
	(1)	(2)
life_built	-0.001*** (0.0001)	-0.001*** (0.0002)
life_renovated	-0.0002 (0.0002)	0.0004 (0.0004)
life_built:life_renovated		-0.00001 (0.00001)
Constant	13.113*** (0.008)	13.103*** (0.010)
Observations	21,613	21,613
R <sup>2</sup>	0.007	0.007
Adjusted R <sup>2</sup>	0.006	0.006
Residual Std. Error	0.525 (df = 21610)	0.525 (df = 21609)
F Statistic	70.822*** (df = 2; 21610)	48.048*** (df = 3; 21609)

Note:

\*p<0.1; \*\*p<0.05; \*\*\*p<0.01

## 8.1 Analyse des résultats

TABLE 59 – Aperçu de la base de données de l'analyse 5 :

	price	life_built	life_renovated
18130	382880	47	13
21026	727000	5	4
7946	237000	56	1
5269	495000	75	9
10116	555000	10	24
5396	299800	38	11
1569	789500	9	17
1407	300000	15	18
2236	220000	41	0
1682	1230000	13	32

TABLE 60 – Comparaison des performances des régressions linéaires de l'analyse 5

Name	Model	AIC	BIC	R2	R2_adjusted	RMSE	Sigma	Performance_Score
model4	lm	33485.84	33525.75	0.007	0.006	0.525	0.525	1.000
model3	lm	33486.33	33518.26	0.007	0.006	0.525	0.525	0.992
model2	lm	615149.79	615189.70	0.003	0.003	366532.952	366566.875	0.022
model1	lm	615153.65	615185.57	0.003	0.003	366582.615	366608.060	0.000

Nous allons analyser le modèle 4.

$$\mathcal{M}_4 : \log(y_i) = \alpha + \beta_{22}x_{i,22} + \beta_{23}x_{i,23} + \beta_{22,23}x_{i,22}x_{i,23} + z_i$$

Pour  $\alpha$ , nous avons moins de 0,1% de chance de se tromper en rejetant l'hypothèse nulle. Par conséquent, nous rejetons l'hypothèse nulle et nous en déduisons que la constante du modèle 4 est différente de 0 au seuil de 0,1%.

Pour  $\beta_{22}$ , nous avons moins de 0,1% de chance de se tromper en rejetant l'hypothèse nulle. Par conséquent, nous rejetons l'hypothèse nulle et nous en déduisons que le coefficient du modèle 4 de la variable explicative `life_built` est différent de 0 au seuil de 0,1%.

Pour  $\beta_{23}$ , nous avons 34,6% de chance de se tromper en rejetant l'hypothèse nulle. Par conséquent, nous conservons l'hypothèse nulle et nous en déduisons que le coefficient du modèle 4 de la variable explicative `life_renovated` est égale à 0 au seuil de 0,1%.

Pour  $\beta_{22,23}$ , nous avons 11,5% de chance de se tromper en rejetant l'hypothèse nulle. Par conséquent, nous coonservons l'hypothèse nulle et nous en déduisons que le coefficient du modèle 4 de la variable explicative `life_built.life_renovated` est égale à 0 au seuil de 0,1%.

Nous avons  $R^2 = 0,007$  dans le modèle 4 donc 0,7% de la variation du prix des logements est expliquée par la surface intérieure de ces derniers.

Pour  $F$ , nous avons moins de 0,1% de chance de se tromper en rejetant l'hypothèse d'absence globale de significativité. Par conséquent, nous rejetons l'hypothèse d'absence globale de significativité et nous en déduisons que le modèle 4 est globalement avéré au seuil de 0,1%.

## 8.2 Analyse des résidus

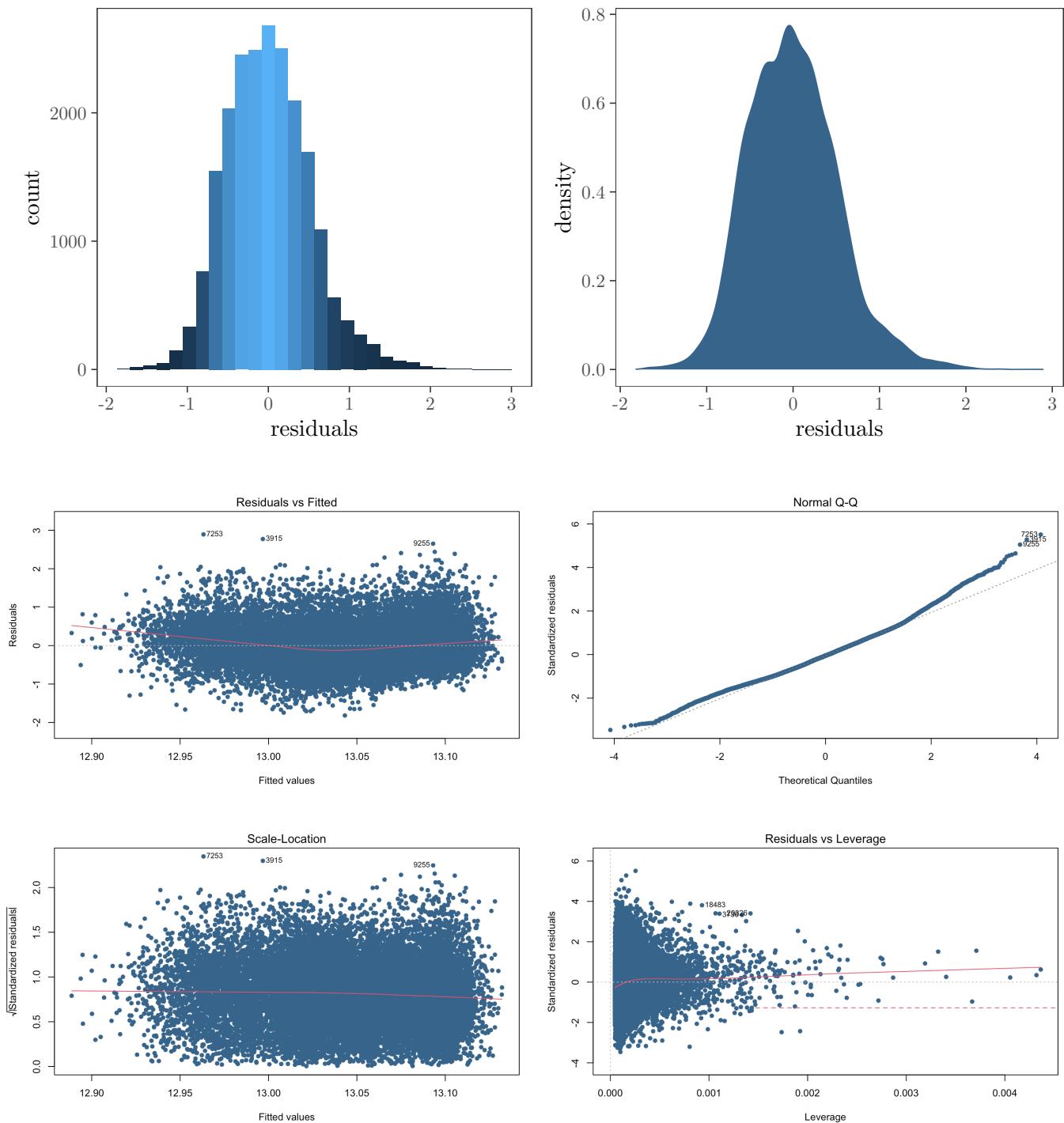


FIGURE 12 – Graphiques de l'analyse des résidus

TABLE 61 – Tests de l’analyse des résidus :

	Harrison_McCabe	Breusch_Pagan	Durbin_Watson
p-value	0.458	0.000	0.005
statistic	0.500	81.297	1.965

Pour  $DW$  nous avons 0,5% de chance de se tromper en rejetant l’hypothèse d’absence d’autocorrélation. Par conséquent, nous conservons l’hypothèse d’absence d’autocorrélation et nous en déduisons que les résidus ne sont pas autocorrélés au seuil de 0,1%.

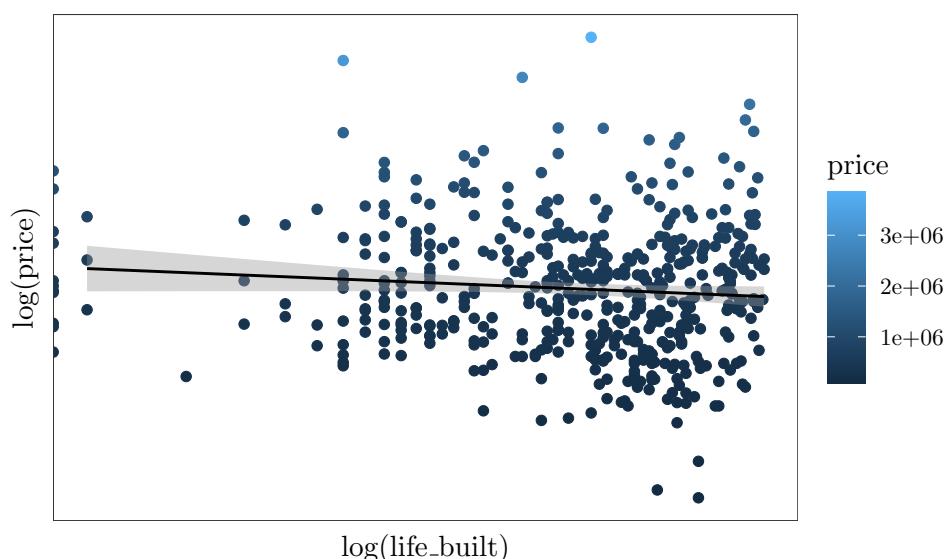
Nous observons que les résidus sont globalement répartis le long de la droite. Néanmoins nous en déduisons que les résidus ne suivent pas une loi normale d’après le graphique représentant la densité des résidus.

Pour  $BP$ , nous avons moins de 0,1% de chance de se tromper en rejetant l’hypothèse d’homoscédasticité. Par conséquent, nous rejetons l’hypothèse d’homoscédasticité et nous en déduisons que les résidus sont hétéroscédastiques au seuil de 0,1%.

Nous observons l’existence de résidus qui ne sont globalement pas linéaires et qui tendent vers la distance de Cook. Nous en déduisons que, dans une moindre mesure, les logement 18483, 9255, 7253, 3915 et 3736 ont un résidu aberrant.

Les hypothèses nécessaires des résultats et des résidus sont avérés au seuil de 0,1%. Notre constante, notre coefficient, notre significativité, notre linéarité des résidus, notre absence d’autocorrélation des résidus et notre appartenance à une loi normale des résidus sont avérés au seuil de 0,1%. L’hypothèse d’hétéroscédasticité ne pose pas de problème car le prix des logements et la variance du prix des logements ne sont pas homogènes.

En conclusion, nous avons 0,1% de chance de se tromper en affirmant que plus les logements sont récents et plus le prix de ces derniers est élevé selon la relation  $\mathcal{M}_4$ .



## 9 $\mathcal{A}_6$ : le prix d'un logement dans la ville de Seattle dans l'État de Washington aux États-Unis

Nous allons segmenter la base de données `BDD_data.csv` en ne prenant en compte que les logements coûtant moins de 645 000 dollars.

```
data.analyse6 <- data[ data$price <= 645000, ]
```

$$\mathcal{A}_6 = \{\mathcal{M}_1\}$$

TABLE 62 – Régressions linéaires de l'analyse 6 :

	<i>Dependent variable:</i>	
	log(price)	
	(1)	(2)
log(sqft_living)	0.829*** (0.010)	0.460*** (0.010)
bedrooms	-0.089*** (0.003)	-0.055*** (0.003)
bathrooms	0.137*** (0.006)	0.089*** (0.006)
floors	0.134*** (0.006)	0.109*** (0.006)
condition2	-0.002 (0.071)	0.064 (0.062)
condition3	0.204*** (0.066)	0.297*** (0.057)
condition4	0.210*** (0.066)	0.290*** (0.057)
condition5	0.259*** (0.066)	0.324*** (0.058)
life_built	0.005*** (0.0001)	0.003*** (0.0001)
Constant	6.160*** (0.091)	8.823*** (0.087)
Observations	21,613	16,240
R <sup>2</sup>	0.532	0.273
Adjusted R <sup>2</sup>	0.532	0.273
Residual Std. Error	0.360 (df = 21603)	0.301 (df = 16230)
F Statistic	2,730.721*** (df = 9; 21603)	678.560*** (df = 9; 16230)

Note:

\*p<0.1; \*\*p<0.05; \*\*\*p<0.01

Les régressions linéaires de l'analyse 6 ne sont pas plus intéressantes que les régressions linéaires des analyses 1, 2, 3, 4 et 5.

## 10 Synthèse

Vous devez vous référer à l'annexe `DevoirEconometrieLubinVialatteAnnexe (2).pdf`.

## 11 tools

Nous nous sommes servis de ces `tools` :

```

tool.table <- function( b, c, d ){
  a <- b %>%
    kable( caption = c,
           digits = 3,
           booktabs = T ) %>%
    kable_styling( full_width = F,
                  position = "center",
                  latex_options = c( "striped",
                                     "condensed",
                                     "hold_position",
                                     d ) )
  return( a )
}# pour les tableaux,
# a = ( b = data, c = caption, d = "scale_down" ),

tool.summary <- function( b ){
  a <- c( mean( b,
                na.rm = T ),
          sd( b,
               na.rm = T ),
          quantile( b,
                     na.rm = T ) )
  a <- round( a,
             3 )
  names( a ) <- c( "Moyenne",
                   "Ecart-type",
                   "Minimum",
                   "Q1",
                   "Q2",
                   "Q3",
                   "Maximum" )
  return( a )
}# pour les résumés des variables,
# a = ( b = data ),

tool.durbin <- function( b ){
  test1 <- dwtest( b )
  a <- c( test1$p.value,
          test1$statistic )
  a <- round( a,
             3 )
  names( a ) <- c( "p.value",
                   "DW" )
  return ( a )
}# pour les tests de Durbin-Watson,
# a = ( b = data ),

tool.shapiro <- function( b ){
  test1 <- shapiro.test( b )
  a <- c( test1$p.value,
          test1$statistic )
  a <- round( a,
             3 )

```

```

names( a ) <- c( "p.value",
                 "W" )
return( a )
}# pour les tests de Shapiro-Wilk,
# a = ( b = data ),

tool.breusch <- function( b ){
  test1 <- bptest( b )
  a <- c( test1$p.value,
          test1$statistic )
  a <- round( a,
              3 )
  names( a ) <- c( "p.value",
                  "BP" )
  return ( a )
}# pour les tests de Breusch-Pagan,
# a = ( b = data ),

tool.harrison <- function( b ){
  test1 <- hmctest( b )
  a <- c( test1$p.value,
          test1$statistic )
  a <- round( a,
              3 )
  names( a ) <- c( "p.value",
                  "HMC" )
  return ( a )
}# pour les tests de Harrison-McCabe,
# a = ( b = data ),

tool.student <- function( b, c ){
  test1 <- var.test( b ~ c )
  test2 <- t.test( b ~ c,
                  equal = test1$p.value > 0.05 )
  a <- c( table( c[ !is.na( b ) ] ),
         ifelse( test2$p.value > 0.05,
                 "Oui",
                 "Non" ),
         round( c( test2$estimate,
                   test2$p.value ),
                3 ) )
  names( a ) <- c( names( table( c[ !is.na(b) ] ) ),
                  "var.equal",
                  names( test2$estimate ),
                  "p.value" )
  return( a )
}# pour les tests de Student de comparaison des moyennes et des variances,
# a = ( b = data1, c = data2 ),

tool.chi2 <- function( b, c ){
  test1 <- chisq.test( b,
                       c )
  a <- c( min( test1$expected ),

```

```

    test1$p.value )
a <- round( a,
            3 )
names( a ) <- c( "Eff_théorique_min",
                  "p-value" )
return( a )
}# pour les tests du Chi 2 d'indépendance des variables,
# a = ( b = data1, c = data2 ),

```

## 12 packages

Nous nous sommes servis de ces packages :

```

library( readxl )# pour la base de données,
library( dplyr )# pour la syntaxe,
library( tidyverse )
library( forcats )# pour les vecteurs,
library( lmtest )# pour les tests,
library( performance )
library( see )# pour le package performance,
library( qqplotr )# pour le package performance,
library( maps )# pour les cartes,
library( mapdata )# pour les cartes,
library( stringr )
library( viridis )

library( plm )# pour les régressions,
library( car )# Pour les régressions,
library( carData )# Pour les régressions,
library( stargazer )# pour les régressions,
library( lmtest )# pour les régressions,
library( statsr )# pour les régressions,
# library( summarytools )# pour les régressions,

library( printr )# pour les tableaux,
library( knitr )# pour les tableaux,
library( kableExtra )# pour les tableaux,
library( modelsummary )# pour les tableaux,
library( gtsummary )# pour les tableaux,

# library( ggplot1 )# pour les graphiques,
library( ggplot2 )# pour les graphiques,
library( ggcrrplot )# pour les graphiques,
library( ggfortify )# pour les graphiques,
library( ggpibr )# pour les graphiques,
library( ggrepel )# pour les graphiques,
library( ggridges )# pour les graphiques,
library( ggsci )# pour les graphiques,
library( ggsignif )# pour les graphiques,

```