

Équipe 2

Travail fait par

Matis Brassard-Verrier (111 182 740)

Alyson Marquis (111 183 605)

Alexis Picard (111 182 200)

Samuel Provencher (111 181 794)

Apprentissage statistique en actuariat

ACT-3114

Rapport 1

Présenté à

Marie-Pier Côté

École d'actuariat
Université Laval
27 février 2020

Table des matières

Introduction	1
Analyse exploratoire des données	2
Traitement des erreurs	2
Analyse univariée	2
Variables explicatives	2
Variable réponse	5
Analyse bivariée	6
Création de variables explicatives	12
Réduction de la dimensionnalité	14
Conclusion	19
Bibliographie	20
Annexe	21

Introduction

Dans le cadre du travail, nous allons tenter de modéliser le prix de vente des maisons dans la région de Seattle (King County, USA) en utilisant de nombreuses caractéristiques ayant une incidence sur la valeur d'une maison. La variable réponse à prédire, soit le prix de vente d'une maison, est une valeur positive évaluée en dollars américains. La modélisation de cette variable pourrait être utile pour différentes raisons dans un contexte actuariel. Comme la somme assurée d'une maison a un lien très fortement proportionnel à son prix de vente, une compagnie d'assurance pourrait être intéressée de modéliser le prix de vente de maisons dans des nouveaux développements immobiliers afin de tenter de prédire les futures soumissions d'assurance habitation et d'offrir des offres personnalisées aux acheteurs de ces nouvelles maisons. Dans un autre contexte, au niveau de la gestion des risques, certains assureurs ont un portefeuille de prêts hypothécaires ou utilisent des produits dérivés sur prêts hypothécaires pour se couvrir du risque (*hedging*). Ainsi, il pourrait être intéressant d'avoir une estimation des montants de prêts hypothécaires dans une région donnée en se basant sur le prix de vente des maisons afin de mieux gérer le risque de la compagnie. La pertinence de trouver cette variable qu'est le prix de vente des maisons devient alors fort intéressante.

Le jeu de données utilisé sera le suivant : kc_house_sales (House sales in King County, USA). Il contient de nombreuses variables explicatives qui seront analysées dans la prochaine section.

Analyse exploratoire des données

Tout d'abord, afin de bien comprendre la base de données choisie, une analyse exploratoire des données est nécessaire. La présente section traite des erreurs décelées dans le jeu de données et fournit des informations pertinentes sur les variables exogènes ainsi que sur la variable réponse sous la forme d'une analyse univariée et bivariée.

Traitement des erreurs

La visualisation des données à l'étude a permis de déceler quelques erreurs. Tout d'abord, 10 observations avaient un nombre de salles de bain égal à 0. Étant donné qu'il est impossible d'avoir une maison sans salle de bain et que ces observations ne représentent qu'un faible pourcentage du jeu de données, il a été convenu de les supprimer. Après avoir enlevé ces 10 observations, il a été remarqué que 6 maisons comptaient 0 chambre. Comme ces maisons en question contenaient toutes un terrain et qu'elles représentaient une faible proportion des données, il a été décidé de les enlever. En outre, une observation avait 33 chambres. En se fiant à l'aire habitable de la maison ainsi qu'aux nombre de salles de bain de cette maison, il a été convenu que le nombre de chambres avait subi une erreur de frappe. Puisqu'il était impossible de déterminer avec certitude le véritable nombre de chambres de la maison, il a été décidé de simplement retirer cette observation du jeu de données.

Comme nous avons un nombre considérable d'entrées dans la base de données (plus de 21000), les conséquences d'enlever 17 entrées de données fautives sont minimales. Si jamais le patron de non-réponse des données manquantes en question n'était pas MCAR, nous savons que le fait d'utiliser uniquement les cas complets comme nous avons fait amènera un certain biais. Nous avons convenu que ce biais serait minime et que nous pouvions l'ignorer. Il a été envisagé d'imputer stochastiquement les données manquantes pour éviter le biais, mais nous avons réalisé que les tests statistiques pour déterminer un patron de non-réponse ne sont pas concluants lorsqu'on a une proportion extrêmement faible de données manquantes (respectivement 10, 6 et 1 observations sur plus de 21000).

Analyse univariée

Variables explicatives

La base de données initiale comptait 20 variables explicatives. Or, certaines de ces variables n'étaient pas pertinentes dans la modélisation du prix de vente des maisons. Ainsi, 3 variables explicatives ont été retirées du jeu de données :

- Le numéro d'identification de la vente (*ID*) a été retiré.
- Le code postal (*zipcode*) a été retiré, car le jeu de données contient d'autres variables plus précises sur la localisation des différentes maisons.
- La variable *sqft_living* a été retirée, car il a été réalisé que cette variable était la somme directe des variables *sqft_basement* et *sqft_above*. Un problème évident de multicollinéarité était donc présent et il a été décidé de retirer complètement la variable *sqft_living*.

De plus, la variable *Date* a été modifiée pour être sous le format suivant : année-jour-mois.

Voici d'ailleurs un tableau qui présente les variables explicatives retenues ainsi qu'une brève description de celles-ci.

Variables explicatives	Descriptions
Date	Date de la vente
Bedrooms	Nombre de chambres
Bathrooms	Nombre de salles de bain (0.5 lorsqu'il n'y pas de douche)

Variables explicatives	Descriptions
Sqft_lot	Superficie du terrain en pieds carrés
Floors	Nombre d'étages
Waterfront	Indique si la maison est située sur le bord de l'eau (1 si oui, 0 autrement)
View	Qualité de la vue extérieure allant de 0 à 4
Condition	Condition de la maison allant de 1 à 5
Grade	Qualité de la construction et de la conception allant de 1 à 13
Sqft_above	Superficie habitable au-dessus du niveau du sol en pieds carrés
Sqft_basement	Superficie habitable du sous-sol en pieds carrés
Yr_built	Année de construction
Yr_renovated	Année de rénovation (0 si jamais rénovée)
Lat	Latitude
Long	Longitude
Sqft_living15	Superficie habitable moyenne en pieds carrés des 15 plus proches voisins
Sqft_lot15	Superficie moyenne du terrain en pieds carrés des 15 plus proches voisins

Comme on peut le voir, cette base de données contient des variables assez intéressantes. En effet, elle est composée de variables numériques, temporelles ainsi que spatiales. Ces dernières nous permettront de représenter les données à l'aide d'une carte de la région de Seattle.

Pour poursuivre, le tableau ci-dessous présente les caractéristiques sommatives des variables explicatives.

Variables explicatives	Minimum	1er quantile	Médiane	Moyenne	3ème quantile	Maximum	Écart-type
bedrooms	1,00	3,00	3,00	3,37	4,00	11,00	0,90
bathrooms	0,50	1,75	2,25	2,12	2,50	8,00	0,77
sqft_living	370,00	1430,00	1910,00	2080,32	2550,00	13540,00	918,11
sqft_lot	520,00	5040,00	7618,00	15099,41	10685,00	1651359,00	41412,64
floors	1,00	1,00	1,50	1,49	2,00	3,50	0,54
waterfront	0,00	0,00	0,00	0,01	0,00	1,00	0,09
view	0,00	0,00	0,00	0,23	0,00	4,00	0,77
condition	1,00	3,00	3,00	3,41	4,00	5,00	0,65
grade	3,00	7,00	7,00	7,66	8,00	13,00	1,17
sqft_above	370,00	1190,00	1560,00	1788,60	2210,00	9410,00	827,76
sqft_basement	0,00	0,00	0,00	291,73	560,00	4820,00	442,67
yr_built	1900,00	1951,00	1975,00	1971,00	1997,00	2015,00	29,38
yr_renovated	0,00	0,00	0,00	84,46	0,00	2015,00	401,82
lat	47,16	47,47	47,57	47,56	47,68	47,78	0,14
long	-122,52	-122,33	-122,23	-122,21	-122,13	-121,32	0,14
sqft_living15	399,00	1490,00	1840,00	1986,62	2360,00	6210,00	685,23
sqft_lot15	651,00	5100,00	7620,00	12758,28	10083,00	871200,00	27274,44

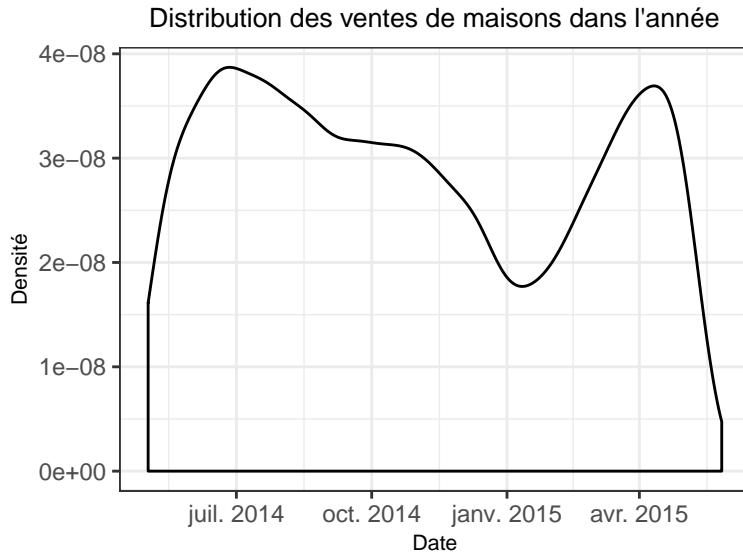
FIGURE 1 – Caractéristiques sommatives des variables explicatives

Il est possible de faire ressortir quelques informations de ce tableau. Tout d'abord, la variable *grade* est une variable qui prend des valeurs entières de 1 à 13. Or, le minimum est de 3. Aucun impact n'est envisagé quant à l'analyse de cette variable puisque les graphiques seront gradués de 3 à 13. De plus, quelques variables ont un très grand écart-type, soit *sqft_lot*, *sqft_lot15*, *sqft_above*, *sqft_basement* et *sqft_living15*. Ainsi, il faudra porter une attention particulière lors de la représentation graphique de ces variables. Une transformation quelconque, telle que la transformation logarithmique, pourrait être de mise. Il est également possible de constater que la proportion de maisons situées sur le bord de l'eau est d'uniquement 1 %.

Maintenant, analysons de plus près certaines de ces variables.

Date

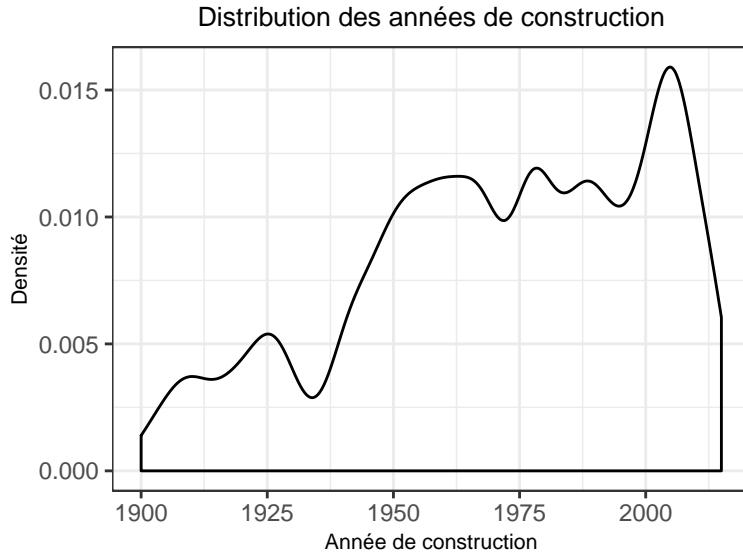
Le graphique suivant présente la densité de la variable *Date*.



Il est possible de constater qu'il semble y avoir un effet de saisonnalité dans la vente de maisons. En effet, l'été semble être une période où il y a beaucoup de ventes de maisons tandis que l'hiver semble être une période où la vente de maisons est moins fréquente. C'est important d'être au fait de cette tendance dans le domaine immobilier.

Yr_built

Le graphique suivant illustre la distribution de l'année de construction.

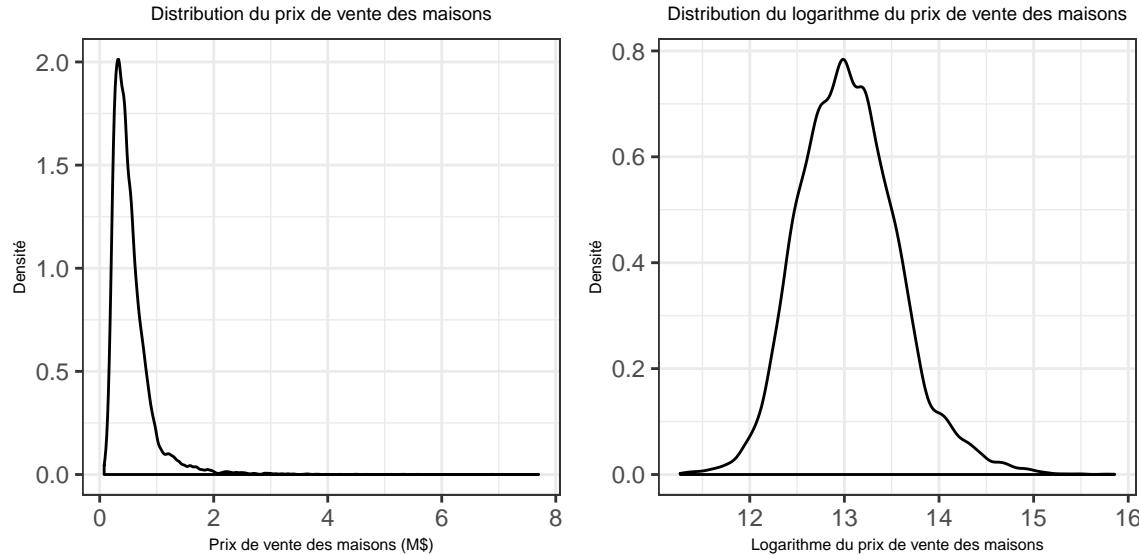


Tout d'abord, il est possible de constater que la base de données contient plus de maisons récentes que de vieilles maisons. On semble remarquer 2 augmentations majeures sur le graphique. En effet, il y a une

hausse vers les années 1950, puis une seconde vers le début des années 2000. De plus, on constate une légère concentration de maisons construites en 1900. Plus précisément, 87 maisons ont été construites en 1900. Or, le nombre de maisons construites en 1901 et 1902 est similaire à celui de 1900. Ainsi, nous avons supposé que les données ont été collectées sur des maisons construites à partir de 1900.

Variable réponse

La variable réponse de notre jeu de données est la variable *price*. Elle représente le prix de vente des maisons dans la région de Seattle et est illustrée par les graphiques suivants.



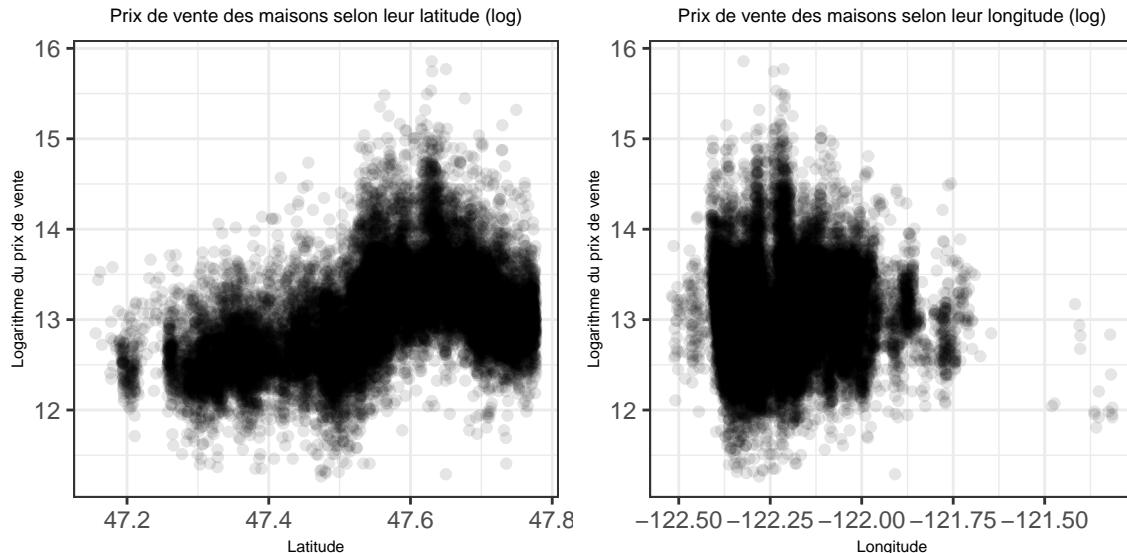
Lorsqu'on regarde le graphique de gauche, où aucune transformation logarithmique n'a été effectuée sur la variable du prix, il est possible de remarquer que la distribution a une forte asymétrie à droite. Cela a pour effet de décaler la distribution à gauche de la médiane et d'étaler la queue de la distribution vers la droite. Ainsi, dans cette représentation graphique, un grand nombre d'observations est regroupé dans des prix plus faibles.

Pour pallier cette asymétrie, la transformation logarithmique a été effectuée sur la variable réponse. Le résultat est présenté sur le graphique de droite. Il est possible de constater que la distribution dans ce graphique est symétrique et qu'elle s'approche de la forme d'une loi normale. Ainsi, il sera plus facile de modéliser le logarithme et d'analyser les résultats.

Analyse bivariée

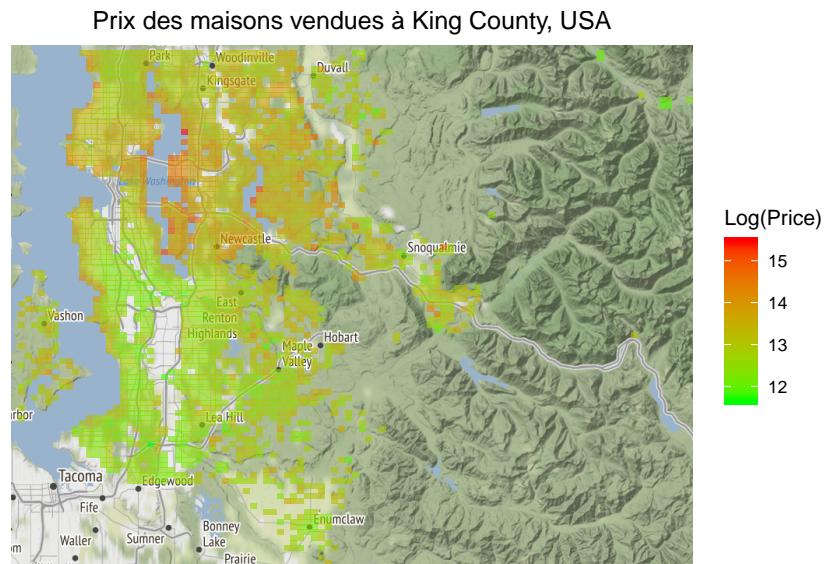
Variables spatiales

Nous pouvons examiner le prix de vente des maisons selon leur latitude et leur longitude avec les graphiques suivants :



Il est assez difficile d'interpréter les tendances dans ces graphiques. Au niveau de la latitude, on voit que les maisons au sud de la région sont peu chères, puis ce prix augmente lorsqu'on monte au nord et qu'on atteint la zone urbaine de Seattle, puis le prix redescend lorsqu'on tombe dans la zone plus rurale au nord de la région. Au niveau de la longitude, il ne semble y avoir aucune tendance intéressante à observer.

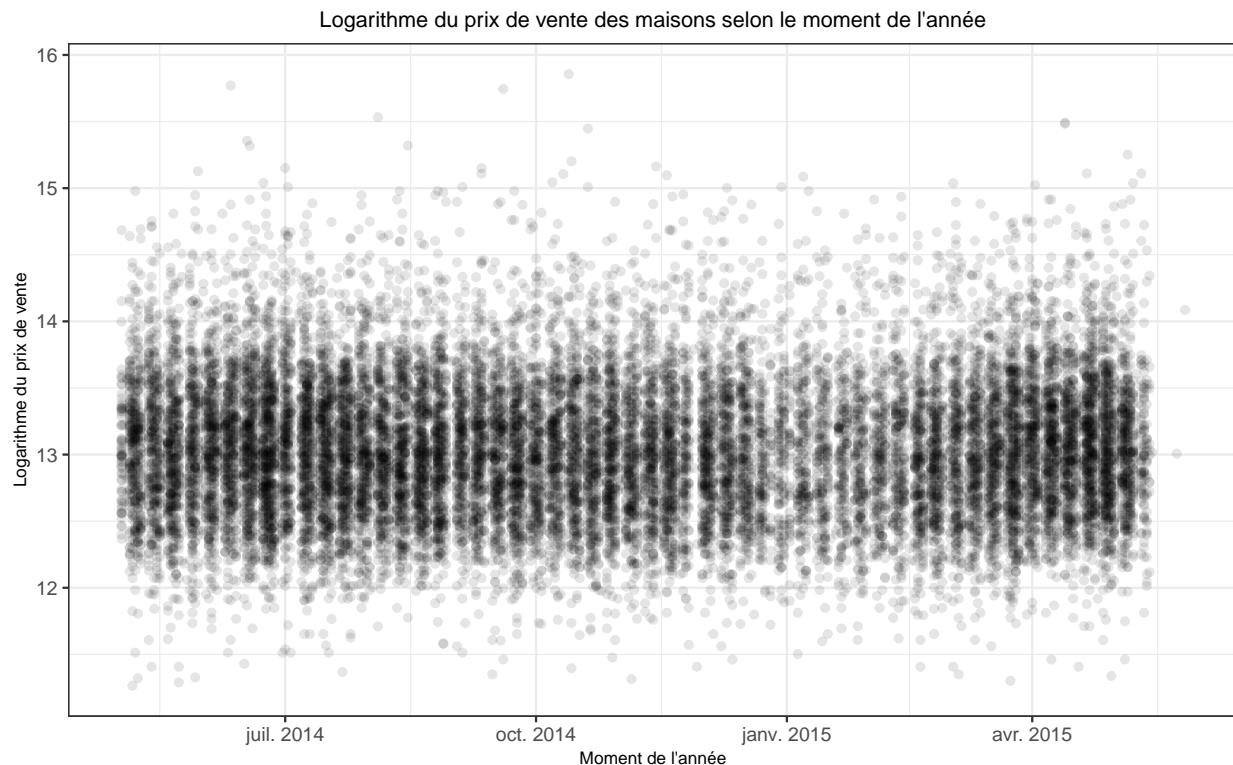
Il a été réalisé que visualiser les prix des maisons de cette manière était peu informatif et que les variables latitude et longitude seraient peu utiles par la suite. Nous avons donc eu comme idée de plutôt visualiser les prix des maisons sous forme de carte thermique de King County :



Il est beaucoup plus facile de faire ressortir des tendances de cette manière. En regardant la carte thermique affichée ci-dessus, il est possible de voir qu'une bonne partie des maisons les plus coûteuses se situent au

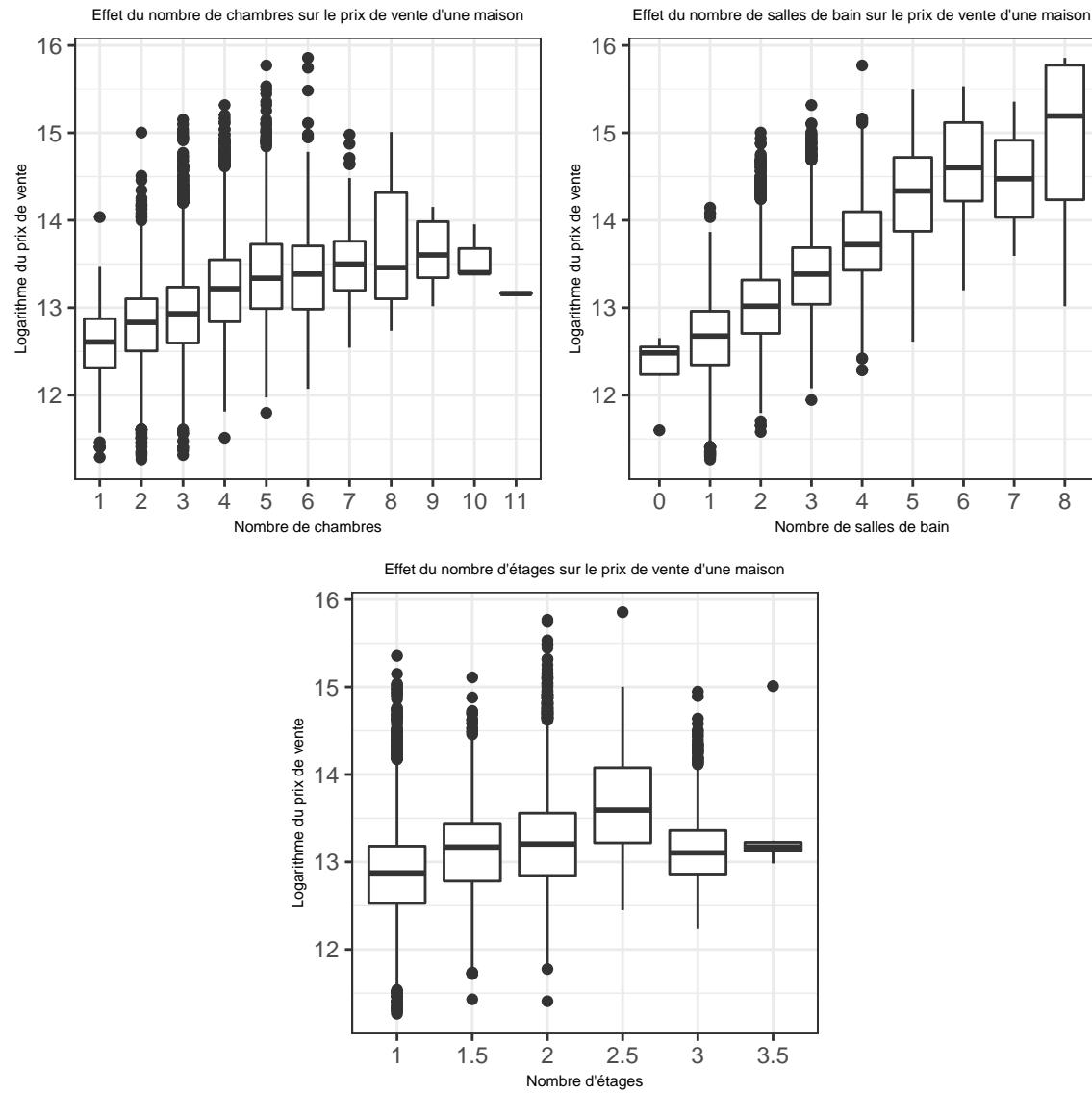
nord-ouest de la région, tout près de l'eau. Cette région semble être la zone urbaine de Seattle. Quant à elles, les maisons situées en région rurale ou en banlieue éloignée ont des prix de vente beaucoup plus bas. Comme la localisation a un impact sur le prix que nous ne voulons pas ignorer, mais que les variables *latitude* et *longitude* ne reflètent pas par elles-mêmes cet impact, nous transformerons ces deux dernières variables en une nouvelle variable plus éloquente à la section “Création de variables explicatives”.

Date en fonction du logarithme du prix



Le graphique ci-dessus montre qu'il n'y a pas de relation linéaire entre le moment de la vente de la maison et le logarithme du prix obtenu. En effet, peu importe la date, le logarithme du prix semble distribué normalement et la variance des observations est constante dans le temps. Il est intéressant de remarquer que les transactions sont fort probablement regroupées par semaine (lundi au vendredi) dû à l'horaire des bureaux de notaires, ce qui veut dire qu'aucune ou peu de transactions sont conclues la fin de semaine.

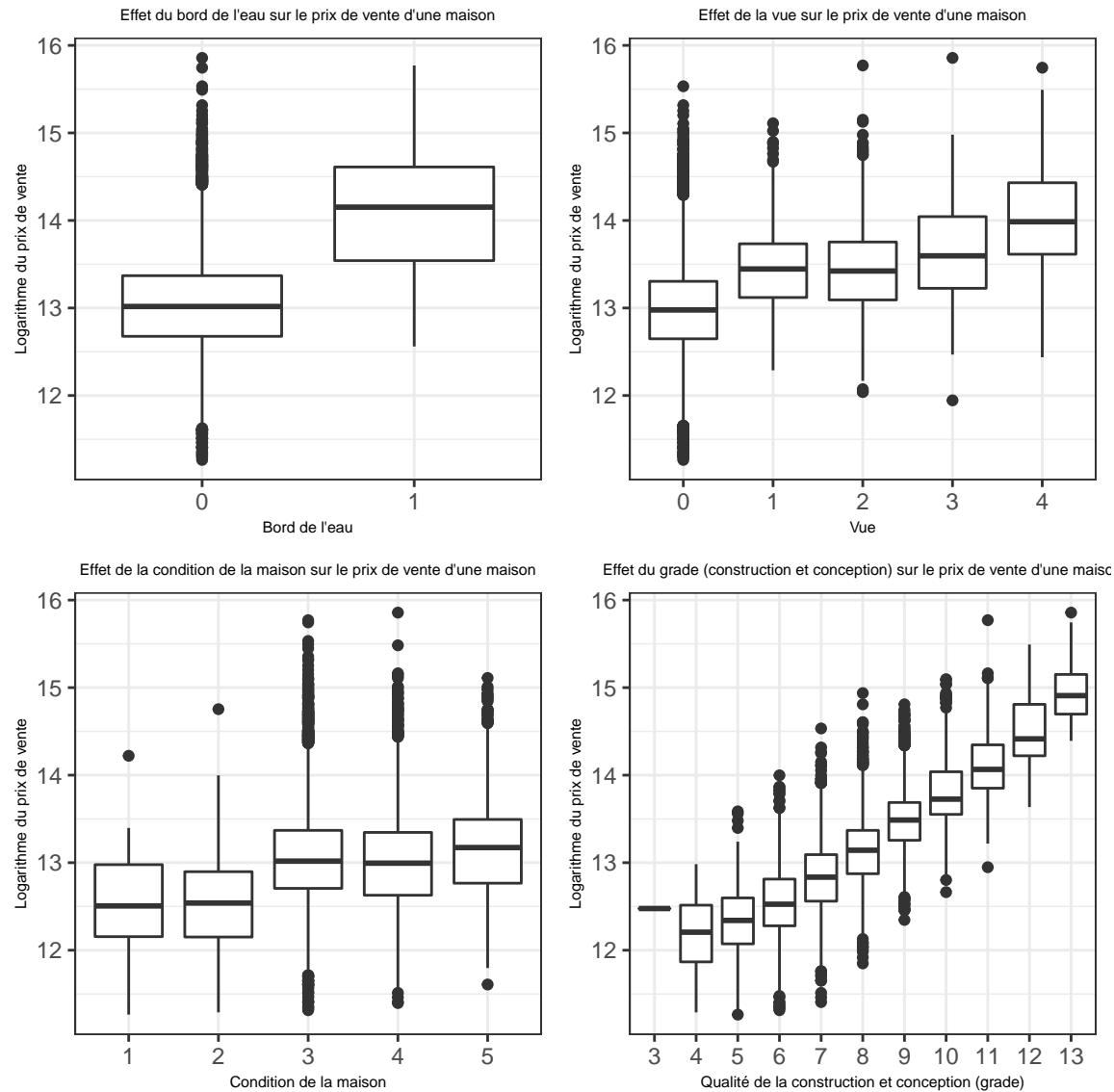
Variables *bedrooms*, *bathrooms* et *floors*



Le constat général de ces trois graphiques est que plus ces variables prennent des valeurs élevées, plus le logarithme du prix augmente. Il s'agit d'un constat assez intuitif, donc ce n'est pas surprenant. Un second constat est qu'il y a une plus grande volatilité pour de faibles valeurs de ces trois variables. En effet, cela est attribuable au fait qu'il y a un grand nombre d'observations pour de faibles valeurs. Pour les valeurs plus à droite dans les graphiques, c'est plus difficile d'analyser, car les cas sont beaucoup plus rares. Le dernier constat à relever est au niveau de la variable *bathrooms*. À l'aide du graphique de cette variable ci-haut, il est possible de constater que l'impact des salles de bains sur la variable réponse est plus important que l'impact des deux autres variables en raison de la pente plus prononcée. Tout porte à croire que ces 3 variables sont reliées à la superficie habitable, car habituellement, les maisons plus dispendieuses sont plus grosses en terme de superficie, de nombre de chambres, étages et salles de bain.

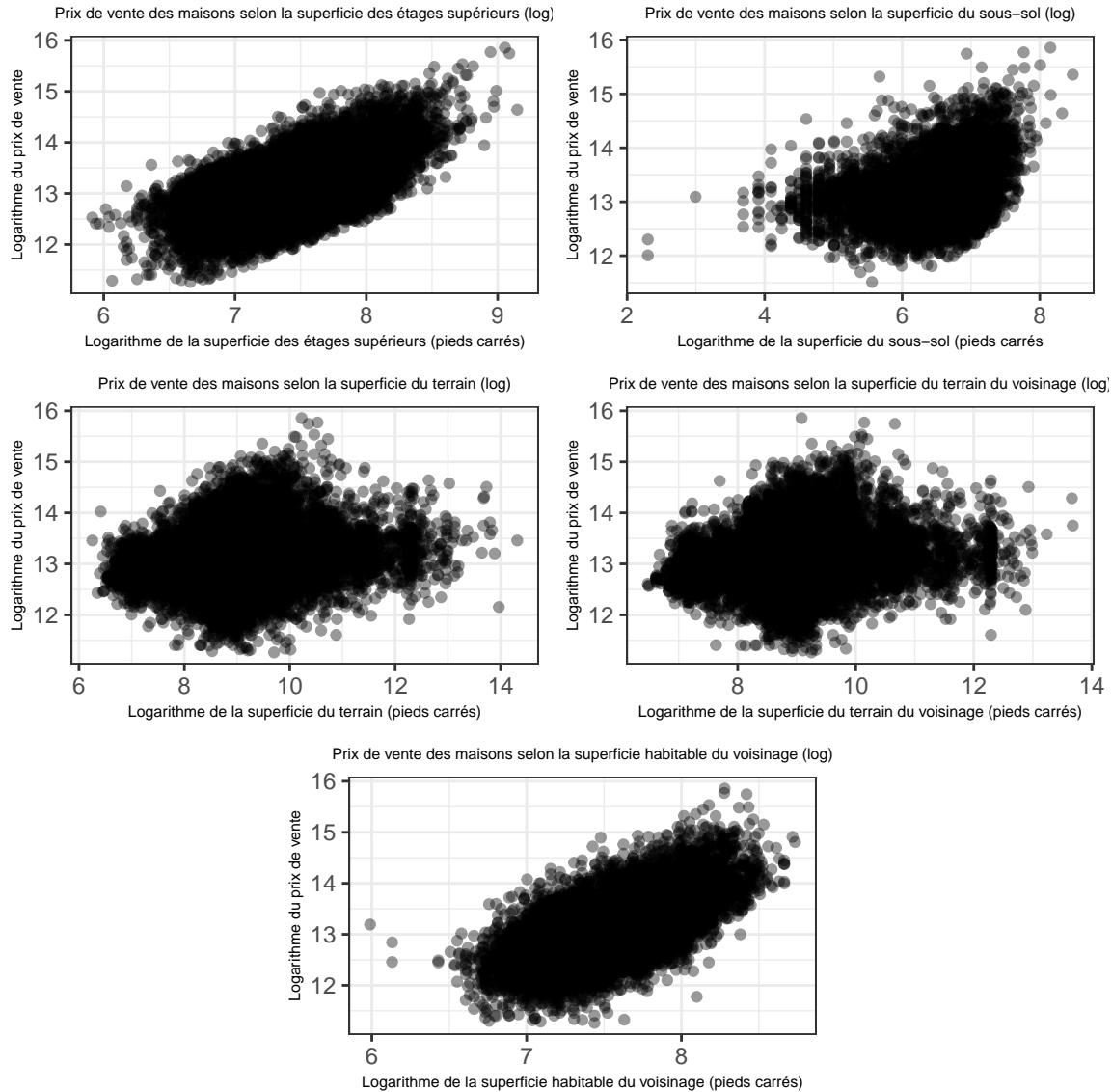
Remarque : il est à noter que pour des fins de visualisation, la variable *bathrooms* a été arrondie vers le bas.

Variables *waterfront*, *view*, *condition* et *grade*



À première vue, il semble que plus ces quatre variables augmentent de niveau, plus le logarithme du prix augmente. Ce constat est logique puisque ce sont toutes des caractéristiques recherchées par les acheteurs. Par contre, au niveau de la variable *condition*, il est plus difficile d'avoir la certitude que la relation tient puisque les boîtes à moustaches sont à des hauteurs statistiquement semblables. Les deux variables dont le changement de niveau amène les plus gros impacts sur la variable réponse sont *waterfront* et *grade*. Pour ce qui est de *waterfront*, il est à noter que seulement 1 % des maisons est situé sur le bord de l'eau, donc pour ces maisons, un prix plus élevé est attendu. En somme, la qualité de la construction et conception et la présence d'un plan d'eau ont un impact significatif sur la valeur d'une propriété dans cette région tandis que la condition de la maison et la vue extérieure ont un impact d'une moindre envergure.

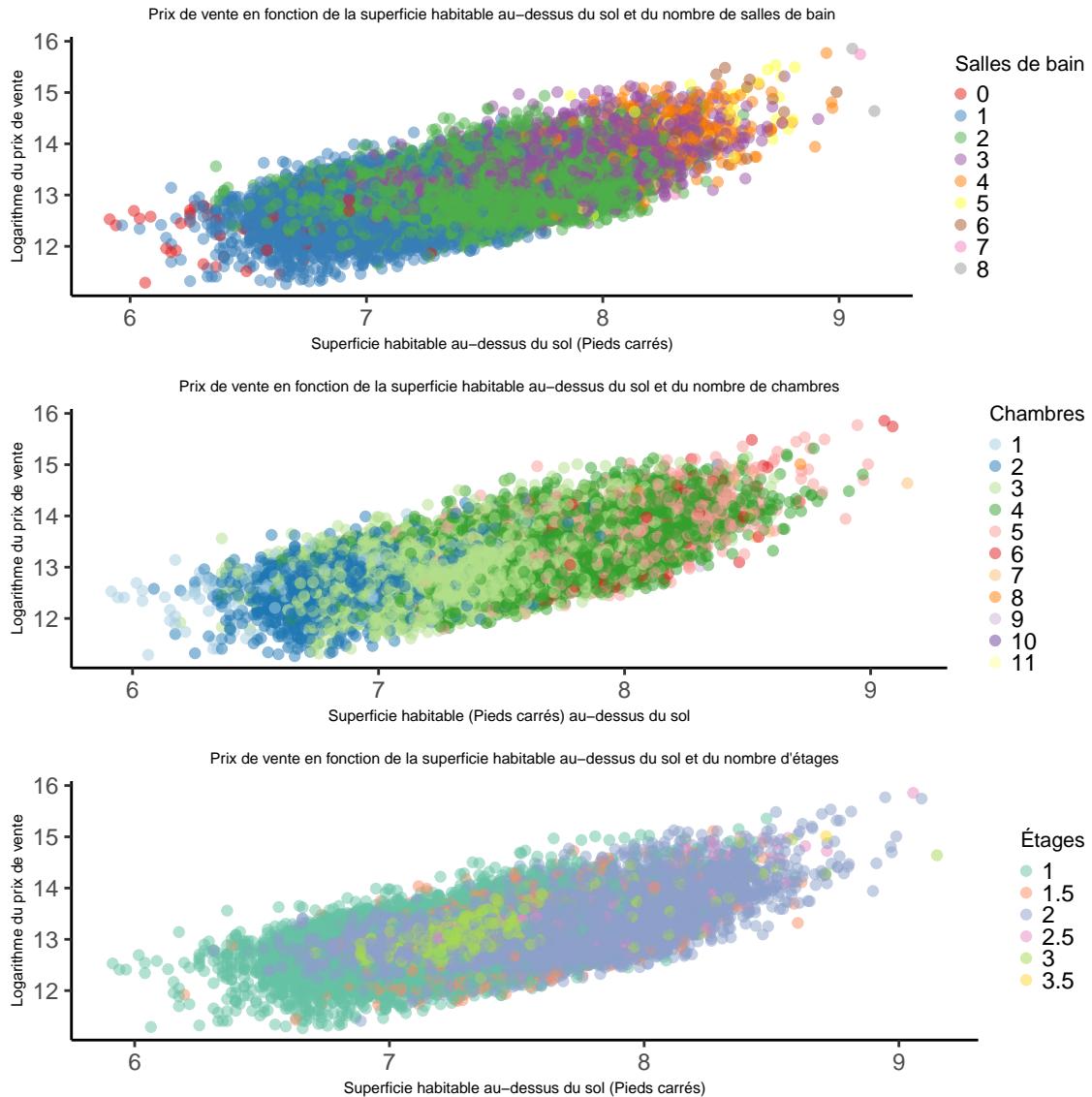
Variables de superficie



Tout d'abord, il a été convenu d'utiliser les données de superficie à l'échelle logarithmique afin de mieux visualiser les données. Ainsi, toute les références aux variables dans les prochaines lignes seront considérées à l'échelle **logarithmique**. En premier lieu, une relation linéaire très évidente est présente entre les variables *sqft_above* ainsi que *sqft_living15* et le prix tel qu'illustré dans le premier et dernier graphique ci-haut. En second lieu, la relation entre la superficie du sous-sol (*sqft_basement*) et le prix, observable sur le second graphique ci-haut, n'est pas aussi évidente. Pour la visualisation, il a été nécessaire de retirer 13110 observations qui n'avaient pas de sous-sol. Pour les maisons possédant un sous-sol, il est possible de conclure que plus le sous-sol est grand, plus le prix de vente sera élevé. Cependant, la relation linéaire n'est pas aussi forte qu'elle l'était pour les deux variables mentionnées plus haut. En troisième lieu, à l'aide des deux graphiques centraux, la superficie du terrain (*sqft_lot*) et la superficie des terrains du voisinage (*sqft_lot15*) ne semblent pas avoir d'impact sur le prix de vente d'une maison. Cela est très surprenant et inattendu, car règle générale, les terrains ont également des valeurs foncières. Une hypothèse pour cette absence de relation est que les maisons du jeu de données sont principalement situées dans une zone urbanisée (région de Seattle) où les terrains seraient de dimensions similaires. Lorsque ces deux variables sont observées de plus près, une grosse masse de données autour de la médiane est découverte, ce qui rend l'hypothèse plausible (se référer au

tableau sommaire de l'analyse univariée). Bref, la superficie totale d'une maison et celles des maisons du voisinage ont un impact significatif sur le prix de vente tandis que la superficie du terrain de la maison et celles des terrains du voisinage ne semblent pas avoir d'impact.

Examinons maintenant plus en détail la relation entre la superficie habitable au-dessus du sol et le prix de vente des maisons selon le nombre de salles de bain, le nombre de chambres et le nombre d'étages.



Tel que mentionné précédemment, il semble avoir une relation entre la variable *sqft_above* et le nombre de chambres, d'étages et salles de bain. Les graphiques ci-haut permettent d'affirmer ce constat. En effet, on remarque que les maisons qui ont plus de chambres, d'étages ou de salles de bain sont généralement les maisons dont la superficie habitable au-dessus du niveau du sol est la plus grande. On observe également une tendance linéaire positive. Autrement dit, les maisons qui ont une plus grande superficie habitable au-dessus du niveau du sol ainsi qu'un grand nombre d'une des 3 variables en question ici se vendent à des prix plus élevés.

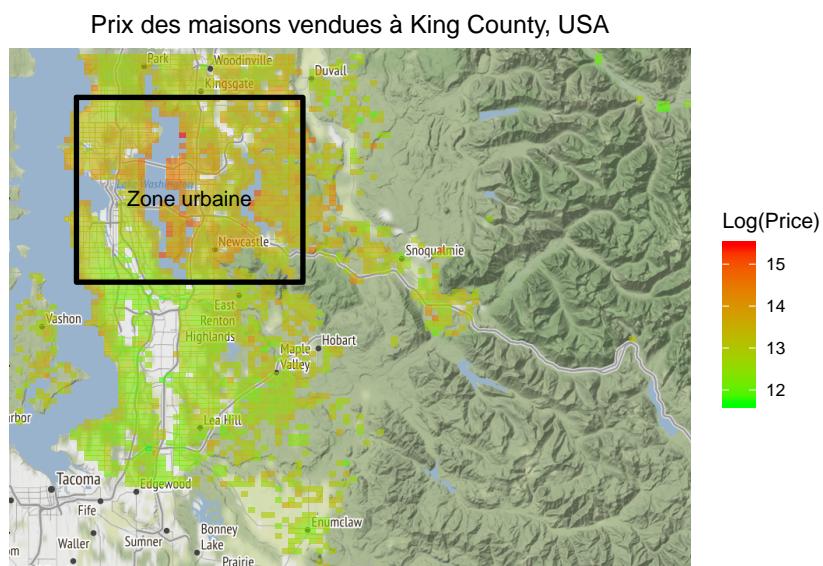
Création de variables explicatives

D'après les données accessibles, il a semblé pertinent d'en utiliser certaines afin de créer d'autres variables explicatives qui seront plus utiles afin de déterminer le prix de vente des maisons de King County.

La première qui a été créée est celle de l'âge de la maison (*age*). Initialement, le jeu de données permettait d'obtenir la date de vente de la maison grâce à la variable *date* et l'année de construction de la maison grâce à la variable *yr_built*. À l'aide de ces deux variables, il est donc facile d'obtenir l'âge de la maison, soit le nombre d'années depuis qu'elle a été construite avant la vente de celle-ci. Analyser l'âge de la maison est plus facile qu'analyser deux dates prises séparément. Combiner ensemble ces deux dates crée donc une variable numérique discrète plus utile pour en déduire le prix d'une maison. De plus, un coefficient de régression associé à l'année de construction aurait été difficile à analyser, puisque l'année la plus ancienne était 1900. Un coefficient de régression sur l'âge sera plus facile à analyser.

La deuxième variable qui a été créée est celle pour savoir si la maison a été rénovée ou non depuis sa construction. Elle a été nommée *reno*. À prime abord, il a été testé si cette nouvelle variable ne devrait pas plutôt être catégorielle ordinaire avec des catégories allant de *10 ans et moins* pour les maisons ayant eu une rénovation dans les 10 années précédant leur vente, *10 ans et plus* pour les maisons ayant eu une rénovation il y a 10 années ou plus et *Jamais rénové* pour les maisons n'ayant jamais été rénovées depuis leur construction. Par contre, après une analyse plus poussée, les catégories autres que celle *Jamais rénové* affichait une moyenne et une médiane de prix de vente similaires. De plus, la proportion de maisons rénovées était faible (4 %). Pour ces deux raisons, il a été décidé que séparer les maisons rénovées en deux catégories semblait exagéré, c'est pourquoi au final, la variable *reno* prend uniquement comme valeurs 1 si la maison a déjà été rénovée depuis sa construction ou 0 si elle n'a jamais été rénovée.

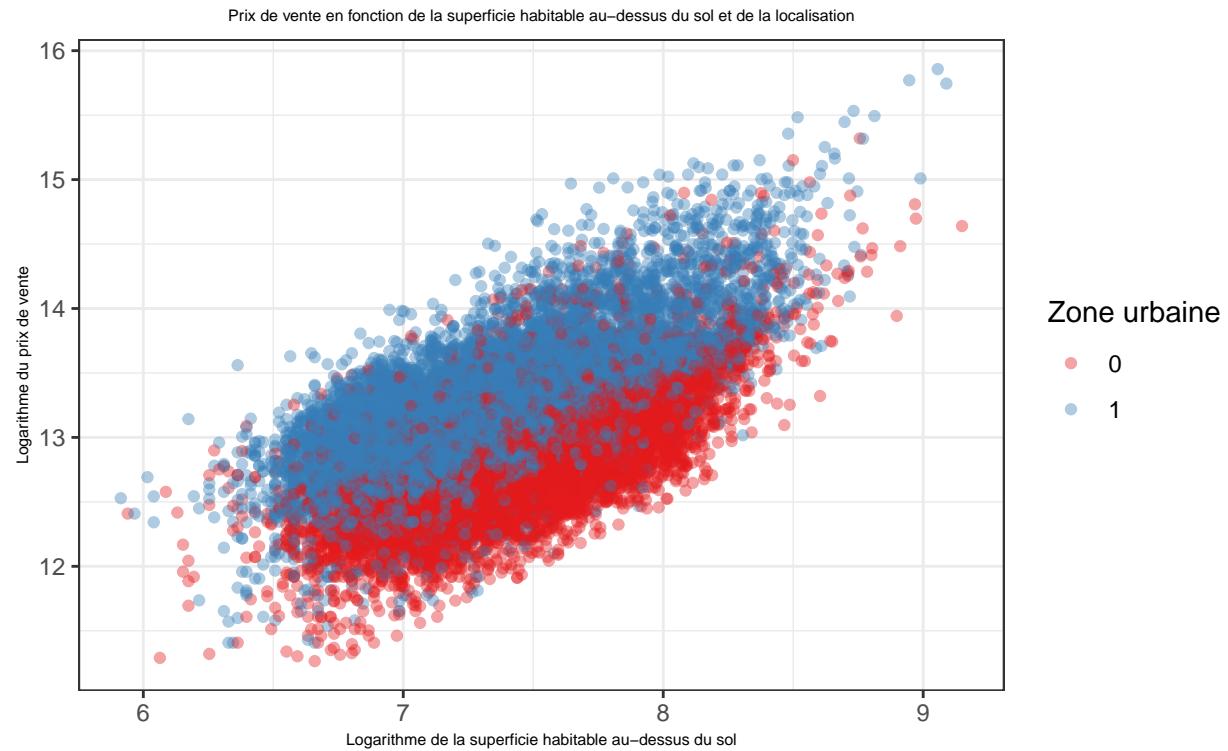
La dernière variable créée est celle représentant la région la plus coûteuse (*expensive_area*). Il a déjà été déterminé à la section "Analyse bivariée" que les variables *latitude* et *longitude* devraient être remplacées par une variable plus éloquente. En examinant la carte thermique dans cette section, il est possible de voir qu'une bonne partie des maisons les plus coûteuses se situe au nord-ouest de la région, tout près de l'eau. Cette région semble être la zone urbaine de Seattle. Nous avons donc décidé de déterminer manuellement un rectangle délimitant cette région la plus coûteuse selon la latitude et la longitude. On peut voir ci-dessous l'ajout de la zone plus coûteuse que nous avons déterminée sur la carte thermique :



La variable *expensive_area* est donc une variable catégorielle qui renvoie 1 lorsque la maison est située dans la zone coûteuse et renvoie 0 lorsque la maison ne se situe pas dans la zone coûteuse.

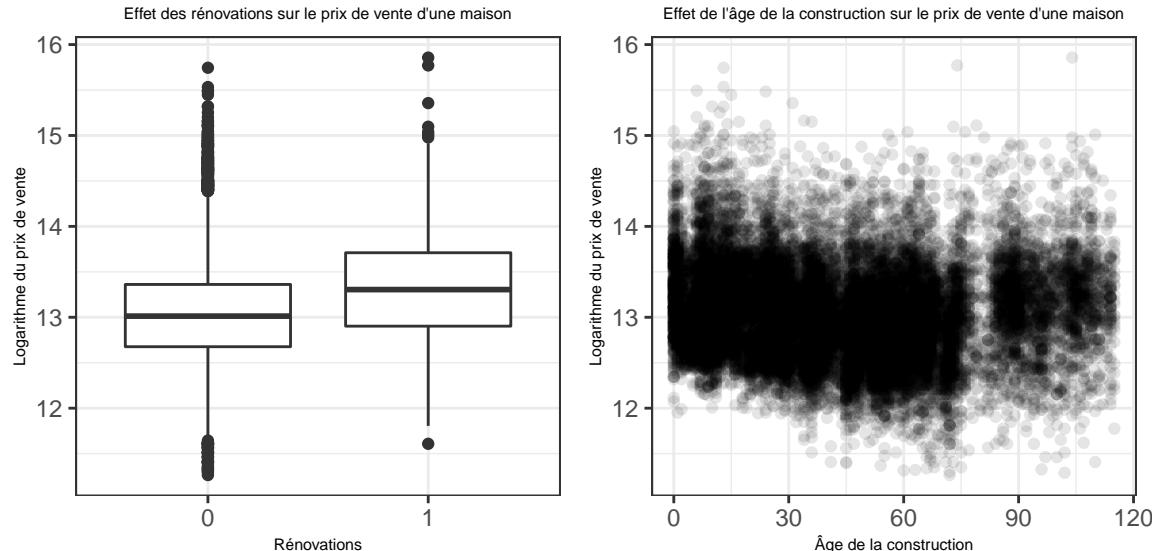
Pour déterminer que la création de cette nouvelle variable est pertinente, examinons si la présence d'une

maison dans la zone sélectionnée a effectivement une influence sur son prix :



On peut bel et bien constater que les maisons localisées dans la zone urbaine ont un prix de vente plus élevé en moyenne.

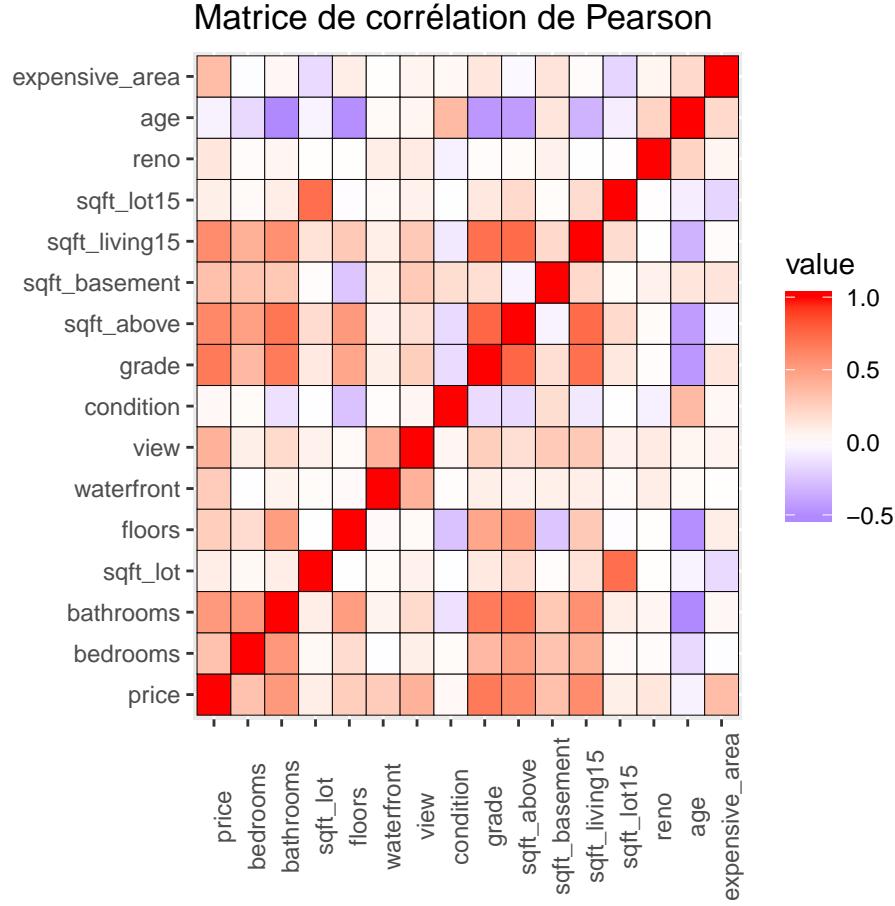
Nous pouvons aussi conduire une analyse bivariée sur les deux autres nouvelles variables explicatives, soit *reno* et *age* :



On peut voir qu'une maisons ayant subi une ou plusieurs rénovations se vendra en moyenne à un prix plus élevé, ce qui est intuitif. Pour ce qui est de l'âge de la maison, aucune tendance nette n'est observée : on peut cependant affirmer que les maisons très récentes (15 ans ou moins) semblent avoir un prix de vente plus élevé que les autres en moyenne, ce qui serait également logique.

Réduction de la dimensionnalité

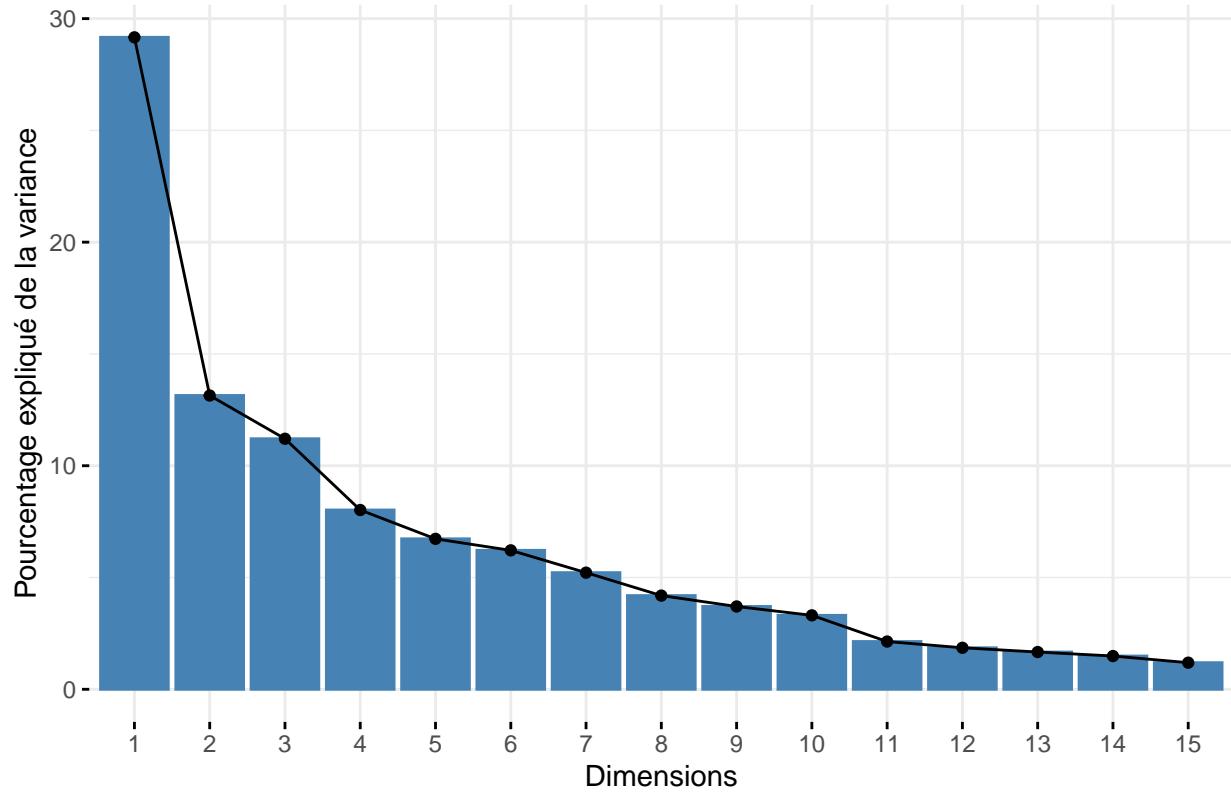
Il se trouve que malgré tout le prétraitement des données et l'analyse univariée et bivariée de celles-ci, on se retrouve avec un jeu de données contenant 16 variables différentes. Ce nombre semble élevé et le nombre de variables fait en sorte qu'on ne peut visualiser efficacement ces données en grande dimension ou même identifier convenablement des maisons exceptionnelles. Afin de voir si certaines variables sont corrélées entre elles, il est possible d'afficher la matrice des corrélations de Pearson :



À l'aide de cette représentation, il est possible de voir les corrélations entre chaque variable du jeu de données. Encore là, il y a beaucoup de couleurs rouge et bleu qui apparaissent, signe de corrélations considérables entre les variables. Cependant, il est difficile de déterminer quelle variable peut correctement compenser pour une autre. C'est pourquoi l'ACP sera appliquée au jeu de données afin de résumer efficacement l'information contenue dans les 16 variables. À partir de nos données qui occupent un certain espace \mathbb{R}^n , l'analyse en composantes principales (ACP) fera une projection vers un nouvel espace \mathbb{R}^m . On cherche naturellement que $m < n$, afin de réduire les besoins en ressources de calcul. Avec cette méthode, il sera plus facile de faire de la visualisation, puisque par exemple, on pourra observer deux composantes de l'ACP à la fois dans un espace \mathbb{R}^2 . Comme chaque composante de l'ACP contiendra des informations venant de plusieurs des variables initiales, cette visualisation permettra donc de ressortir des tendances dans les données qui tiennent compte de plusieurs variables à la fois.

Ici, l'espace initial est \mathbb{R}^{16} . L'ACP nous permettra de réduire le nombre de variables résumant les caractéristiques de chacune d'entre elles, en minimisant la perte de variance et donc d'information. Tout d'abord, nous allons tenter de déterminer le nombre optimal de dimensions à conserver à l'aide d'un diagramme d'éboulis et de l'analyse des vecteurs propres résultant de l'ACP.

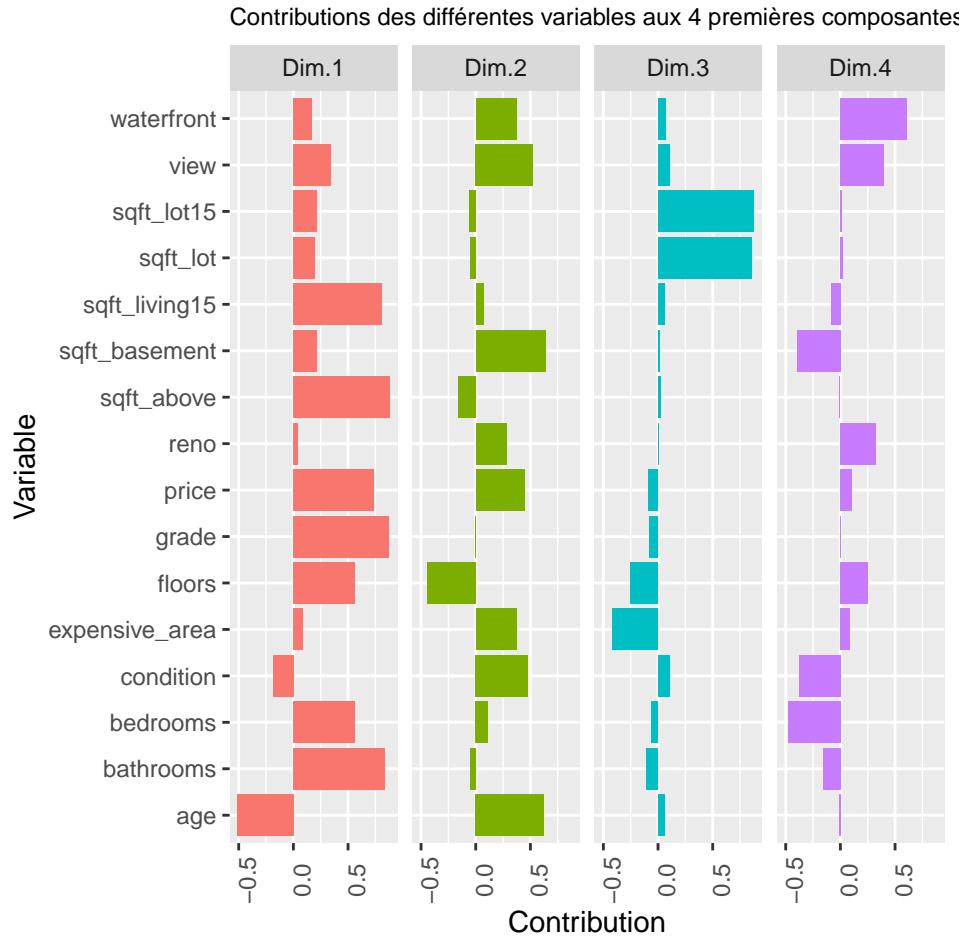
Diagramme d'éboulis



	Valeur propre	Pourcentage de la variance	Pourcentage cumulatif de la variance
comp 1	4.6659972	29.162482	29.16248
comp 2	2.1025407	13.140879	42.30336
comp 3	1.7930991	11.206869	53.51023
comp 4	1.2831613	8.019758	61.52999
comp 5	1.0767224	6.729515	68.25951
comp 6	0.9944390	6.215244	74.47475
comp 7	0.8348115	5.217572	79.69232
comp 8	0.6706555	4.191597	83.88392
comp 9	0.5927129	3.704456	87.58837
comp 10	0.5293085	3.308178	90.89655
comp 11	0.3417558	2.135974	93.03252
comp 12	0.2971890	1.857431	94.88996
comp 13	0.2666628	1.666643	96.55660
comp 14	0.2380467	1.487792	98.04439
comp 15	0.1905277	1.190798	99.23519
comp 16	0.1223698	0.764811	100.00000

En utilisant la méthode du coude en se référant au diagramme d'éboulis, des choix sensés du nombre de dimensions à garder seraient 2, 4, 8 ou 11 dimensions. Cependant, dans le contexte de l'ACP il est intéressant de garder une proportion considérable de la variance des données originales. Garder 2 ou 4 dimensions résulterait respectivement en 42 % ou 62 % de la variance originale, ce qui n'est pas considéré comme étant assez. Cependant, avec 8 dimensions, 84 % de la variance originale est expliquée, ce qui dépasse le seuil de 80 %. Nous souhaitons aussi avoir une réduction considérable du nombre de dimensions, donc le dernier choix possible de 11 nous semble trop élevé. Nous allons donc conserver 8 composantes principales pour la suite.

Examinons maintenant les contributions des variables initiales sur les 4 premières dimensions que nous avons conservées avec l'ACP.



Sur le précédent graphique, il est possible de voir la contribution des variables aux 4 premières composantes créées. Commençons par analyser les deux premières composantes.

On remarque que la première composante semble indiquer la présence de grosses maisons qui viennent juste d'être construites. En effet, la première composante prendra une grande valeur positive lorsque les variables *sqft_living15*, *sqft_above*, *floors*, *bedrooms* et *bathrooms* seront élevées. Toutes ces variables sont de parfaits indicateurs quand à la grosseur de la maison. Il n'est pas rare qu'une grosse maison possède ces caractéristiques. De plus, quand la maison est grande et spacieuse, celle-ci coûte de plus en plus cher et a une tendance à être très belle, d'où les 2 variables *price* et *grade* qui apportent une bonne contribution aussi. Vu la contribution de la variable *age* vers le négatif au contraire des autres, il faudrait assumer que pour une valeur élevée de la composante 1, l'âge prendra une faible valeur et donc que la maison sera très récente.

À l'inverse, une petite valeur de la composante 1 indiquerait l'effet contraire de ce qui a été dit précédemment, soit une petite maison ayant été construite depuis longtemps, puisque la variable *age* contribue beaucoup dans les négatifs et les autres valeurs mentionnées ont quant à elles plus d'importance dans le positif.

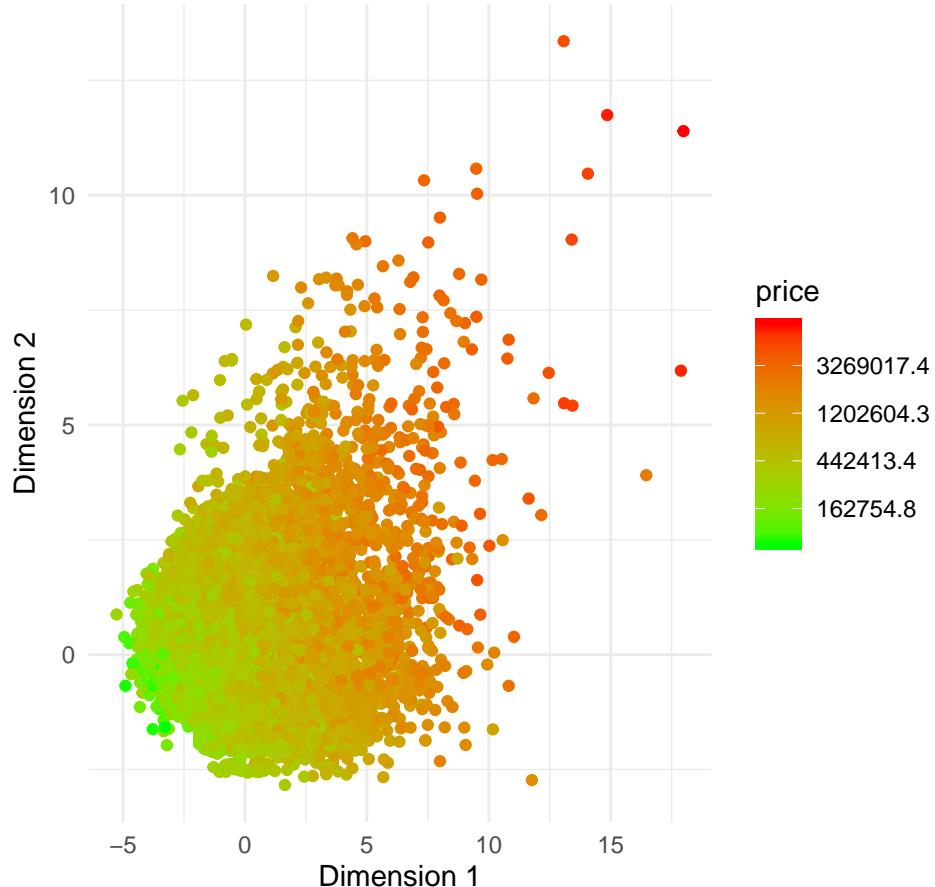
La 2e composante quant à elle semble plutôt indiquer la présence de vieilles maison situées sur le bord de l'eau. Pour une valeur positive élevée de la composante, la contribution viendra majoritairement des variables *view*, *waterfront*, *age* et *sqft_basement*. Les 2 premières indiquent la présence d'une belle vue souvent associée à une vue sur la mer, tandis que l'âge indiquera que la maison est vieille. Si on va plus loin, ceci est logique puisque connaissant bien notre histoire américaine, à l'époque les cours d'eau étaient favorisés vu l'importance des transports maritimes ou même le bonheur d'avoir accès à l'eau pour se baigner ou faire du bateau.

L'importance de l'eau n'a pas diminué pour autant avec le temps et c'est pourquoi des maisons situées près de l'eau sont souvent favorisées et sont donc très chères au final, surtout si elles sont bien situées, d'où la bonne contribution de *condition*, *price* et *expensive area*. À noter l'importance de *sqft_basement* en positif et *floors* en négatif qui seraient indicateurs que les maisons associées par la composante 2 aient soit un gros étage dans les valeurs positives ou plusieurs petits étages dans les valeurs négatives.

Ce qui permettrait de dire qu'une valeur négative de la composante 2 indiquerait plutôt de petites maisons avec beaucoup d'étages, mais non situées près de l'eau.

En allant plus loin avec ces dimensions, il est même possible de confirmer nos affirmations avec les prix de vente de chacune des maisons de notre jeu de données en fonction des 2 premières composantes :

Prix de vente des maisons selon les 2 premières composantes



En effet, le long de l'axe des abscisses, on peut voir le changement de couleur qui passe de vert pâle à gauche à rouge foncé vers la droite, signe que le prix des maisons augmente au fur et à mesure que la composante 1 prend de l'importance. Les grosses maisons récentes ont été associées à de grandes valeurs de cette composante et les petites maisons vieilles à des petites valeurs. Il est donc logique de voir la différence de prix comme décrite précédemment sans argumenter.

Pour la 2e composante, il avait été dit qu'une valeur élevée représentait une vieille maison sur le bord de l'eau avec un étage tandis qu'une valeur faible représentait une maison récente à beaucoup d'étages mais non située près de l'eau. En fait, les caractéristiques positives de ces deux descriptions viennent compenser les négatives et ce, du côté des deux extrêmes. C'est pourquoi on ne peut apercevoir de distinctions précises dans le prix avec la couleur selon les maisons si on regarde à la verticale. Une maison près de l'eau, mais ayant peu d'étages et étant vieille représenterait donc le même prix qu'une maison récente avec beaucoup d'étages, mais qui perd de la valeur vu qu'elle n'est pas près de l'eau.

Enchaînons avec la troisième composante. Cette dernière est particulièrement influencée par les variables *sqft_lot*, *sqft_lot15*, *expensive_area* et *floors*. Une maison ayant une troisième composante de grande valeur possède donc un grand terrain, ses voisins ont également un grand terrain, elle n'est pas située dans la zone métropolitaine et (à un moindre mesure) possède peu d'étages. Selon toute logique, cette dimension représente donc l'état urbain ou rural des maisons. En effet, une maison avec une troisième composante élevée serait une maison en campagne qui possède un très grand terrain et peu d'étages (les maisons de campagne sont souvent plus anciennes et possèdent moins d'étages que les maisons récentes). À l'inverse, une maison située en ville et possédant un petit terrain et beaucoup d'étages aura une valeur très faible pour la troisième composante.

Finalement, la quatrième composante semble représenter les petites maisons en condition moyenne situées sur le bord de mer et possédant une belle vue. En effet, les variables *waterfront* et *view* ont un impact positif sur la composante, alors que les variables *bedrooms*, *condition* et *sqft_basement* ont un impact négatif. On pourrait donc croire que ce sont des petits chalets sur le bord de l'eau qui possèdent peu de chambres, pas de sous-sol et qui sont en condition moyenne, comme des "camps de pêche". À l'inverse, une maison qui ne correspond pas à ce descriptif aura une valeur négative pour la quatrième composante.

Nous allons illustrer les valeurs de la quatrième composante sur la carte de King County pour confirmer notre intuition :

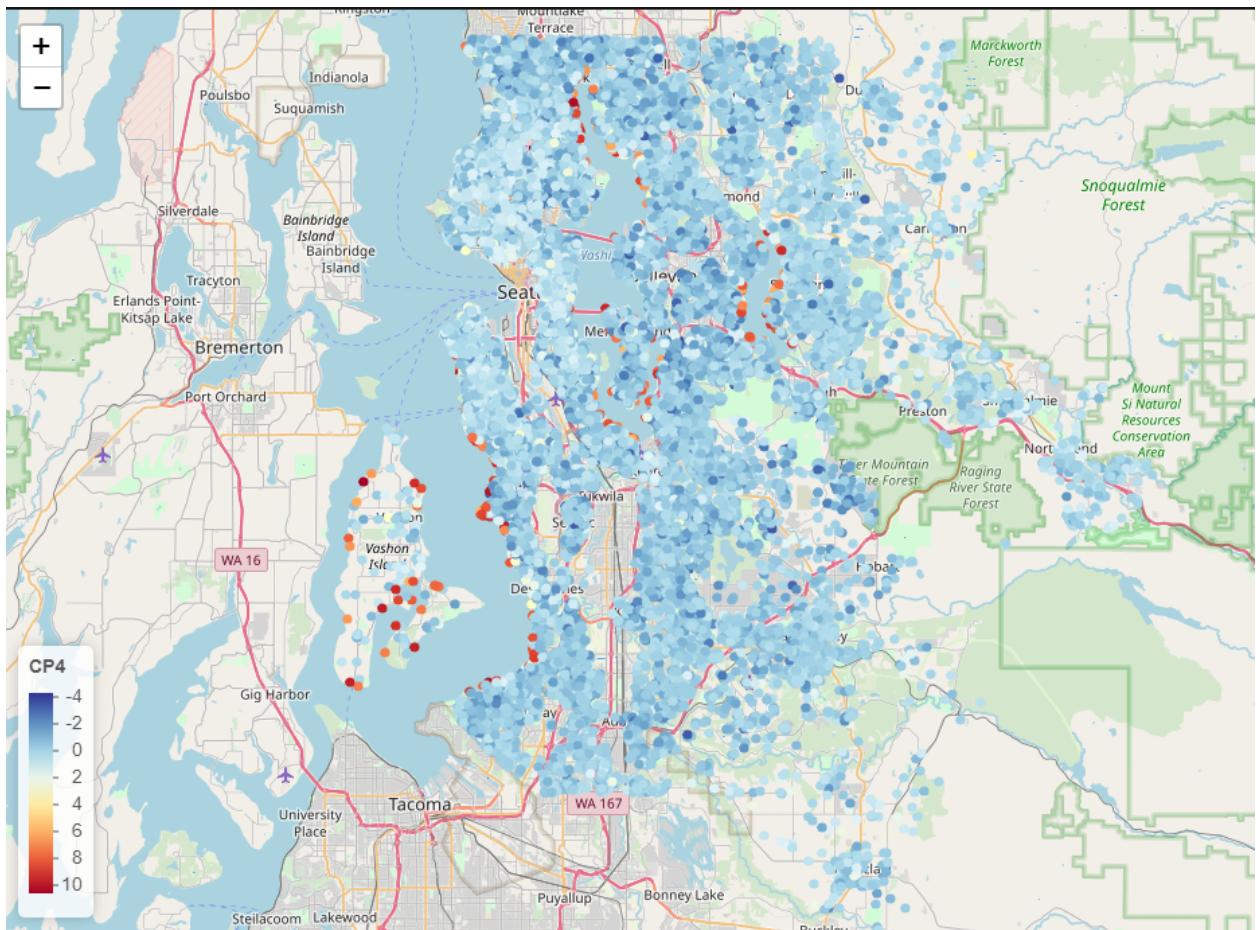


FIGURE 2 – 4ème composante sur la carte de King County

Conclusion

En somme, pour tenter de modéliser le prix de vente des maisons de King County selon les caractéristiques observées des maisons, nous avons tout d'abord trouvé un jeu de données approprié contenant les prix de vente de plus de 21000 maisons entre mai 2014 et mai 2015.

Ensuite, une analyse exploratoire univariée et bivariée a été conduite sur ce jeu de données pour en faire ressortir des tendances. Entre autres, nous avons observé que les caractéristiques physiques des maisons, telles que leur nombre de chambres, d'étages et de salles de bain, la qualité de la construction et la superficie de la maison ont une forte influence positive sur le prix de vente de la maison. C'est aussi le cas de la localisation de la maison (en général, les maisons en ville, sur le bord de l'eau et/ou qui possèdent une belle vue sont plus chères). Certaines caractéristiques avaient quant à elle peu d'impact sur le prix, comme le moment de l'année où la maison a été vendue ou la superficie du terrain.

L'analyse exploratoire nous a également permis de détecter des valeurs aberrantes, qui ont été corrigées. Entre autres, nous avons enlevé des maisons avec aucune chambre, avec aucune salle de bain ou avec un nombre excessif de chambres.

Par la suite, de nouvelles variables explicatives ont été créées lorsqu'il a été jugé que certaines variables n'étaient pas dans une forme assez éloquente. L'une d'entre elles est l'âge de la maison, déterminé à partir de l'année où celle-ci a été construite. Nous avons également créé des variables binaires indiquant :

- Si la maison est située en ville ou non ;
- Si la maison a été rénovée ou non.

Enfin, une analyse en composantes partielles (ACP) a été conduite sur les données pour en réduire la dimensionnalité. L'ACP nous a également permis de mieux visualiser les données à l'aide des nouvelles composantes, qui reflètent les valeurs de plusieurs des variables explicatives d'origine à la fois.

Tout cela nous a permis de mieux comprendre les données en notre possession. La prochaine étape sera de créer un véritable modèle pour tenter de déterminer le prix de vente des maisons selon leurs caractéristiques. Quelques pistes nous semblent intéressantes pour les modèles que nous pourrons tenter :

- Un modèle de régression linéaire multiple semble tout approprié pour le problème que nous essayons ici de résoudre.
- Si une analyse des variables que nous allons utiliser semble dévoiler un problème de multicollinéarité, la régression régularisée ridge pourrait être intéressante.
- Si nous nous rendons compte que seules certaines variables ont un très grand impact sur le prix de vente des maisons, nous pourrions conduire une régression régularisée lasso pour procéder à la sélection des variables.
- Si nous détectons un problème important d'hétéroscédasticité dans les données, une régression pondérée pourrait être envisagée.
- Enfin, les modèles de régression Poisson, binomiale négative ou logistique ne semblent pas avoir un potentiel intéressant pour la suite des choses.

D'autres modèles que nous verrons plus tard dans le cadre du cours, tels que les arbres de régression, les forêts aléatoires et les modèles de *gradient boosting* pourraient également se révéler être intéressants.

Bibliographie

1. Kaggle (2017). House sales in King County, USA. Récupéré de <https://www.kaggle.com/harlfoxem/housesalesprediction>.

Annexe

— **Le nom du jeu de données**

kc_house_sales (House sales in King County, USA)

— **La source**

<https://www.kaggle.com/harlfoxem/housesalesprediction>

— **Une brève description des données (environ deux phrases)**

Cet ensemble de données contient les prix de vente des maisons pour « King County », qui comprend Seattle. Il comprend les maisons vendues entre mai 2014 et mai 2015.

— **La variable réponse et son type**

« House sales » (prix de vente des maisons) - Variable numérique continue

— **La mesure d'exposition (s'il n'y en a pas, le mentionner)**

Il n'y en a pas

— **Cinq variables explicatives et leur type**

Bedrooms : Nombre de chambres – Variable numérique discrète

Bathrooms : Nombre de salles de bain (0.5 est une toilette sans douche) - Variable numérique continue

sqft_living : Superficie de l'espace de vie en pieds carrés - Variable numérique discrète

sqft_lot : Superficie du terrain en pieds carrés - Variable numérique discrète

floors : Nombre d'étages - Variable numérique continue

et plusieurs autres variables bien sur !

— **La taille du jeu de données (nombre d'observations et de variables)**

21 613 lignes pour 21 colonnes (variables)