

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	2
Analyse exploratoire des données	3
Traitement des erreurs	3
Analyse univariée	3
Variables explicatives	3
Variable réponse	6
Analyse bivariée	6
Heatmap	7
Date en fonction du logarithme du prix	7
Variables bedrooms, bathrooms et floors	8
Variables waterfront, view, condition et grade	9
Variables de superficie	10
Variables age, reno et expensive area	10
Création de variables explicatives	12
Réduction de la dimensionnalité	14
Conclusion	17
Bibliographie	18
Annexe	19

Introduction

Dans le cadre du travail, le prix de vente des maisons dans la région de Seattle (King County, USA) sera modélisé en utilisant de nombreuses caractéristiques ayant une incidence sur la valeur d'une maison. Le prix de vente d'une maison est une valeur positive évaluée en dollars américains. Cette valeur modélisée pourrait être utile pour différentes raisons. 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. **Biographie pour la source**

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 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 salle de bain égal à 0. Étant donné qu'il est impossible d'avoir une maison sans salle de bain et que ces observations représentent qu'un faible pourcentage du jeu de données, il a été convenu de supprimer ces 10 observations. Après avoir enlevé ces 10 observations, il a été remarqué que 6 maisons comptaient 0 chambre. En analysant de plus près ces cas, il a été possible de constater que toutes les autres colonnes étaient remplies, donc il ne s'agit pas de données manquantes. De plus, comme ces données 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. En effet, comme nous avons un nombre considérable d'entrées dans la base de données (plus de 21000), les conséquences d'enlever une seule entrée de donnée fautive sont minimes. Il a aussi été envisagé d'imputer stochastiquement la donnée manquante, mais il a été réalisé que les tests statistiques pour déterminer un patron de non-réponse ne sont pas concluants lorsqu'on a uniquement une observation manquante sur plus de 21000 observations.

Analyse univariée

Variables explicatives

La base de données initiales comptaient 20 variables explicatives. Or, certaines de ces variables n'étaient pas pertinentes dans la modélisation du prix de vente des maisons. Ainsi, 2 variables explicatives ont été retirées du jeu de données, soit le numéro d'identification de la vente (*ID*) et le code postal (*zipcode*). Le code postal a été retiré, car le jeu de données contient d'autres variables plus précises sur la localisation des différentes maisons. De plus, la variable *Date* a été mise 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	Nombres de chambres
Bathrooms	Nombres de salles de bain (0.5 lorsqu'il n'y pas de douche)
Sqft_living	Superficie habitable en pieds carrés
Sqft_lot	Superficie du terrain en pieds carrés
Floors	Nombre d'étages
Waterfront	Indique si la maison a une vue sur 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)

Variables explicatives		Descriptions	
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 données contient des variables assez intéressantes. En effet, elle est composée de variables numériques, temporelles ainsi que spatiale. Cela nous permettra de représenter les données à l'aide d'une carte de la région de Seattle.

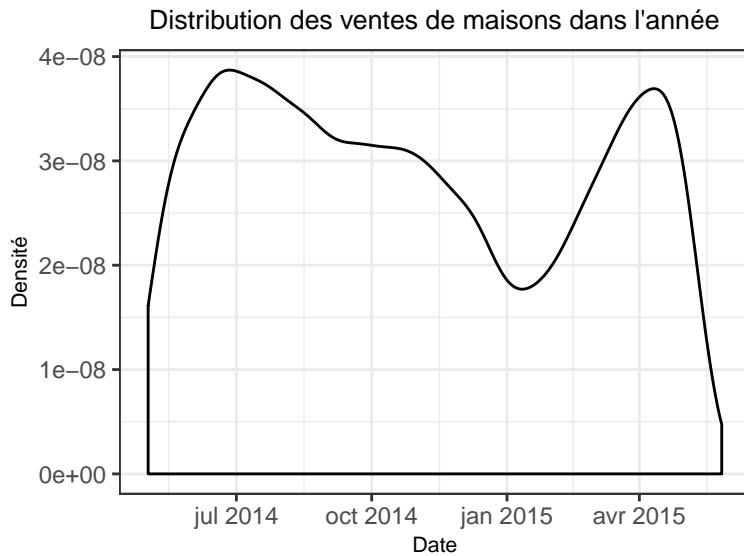
Pour poursuivre, le tableau ci-dessous présente les caractéristiques sommatives des variables explicatives.

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

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. **Rajouter plus de chose sur ça, genre pour simplifier, on enlève 1 et 2 même si ça les trace pas.** De plus, quelques variables ont un très grand écart-type, soit *sqft_lot*, *sqft_above*, *sqft_basement*, *sqft_living15* et *sqft_living*. 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.

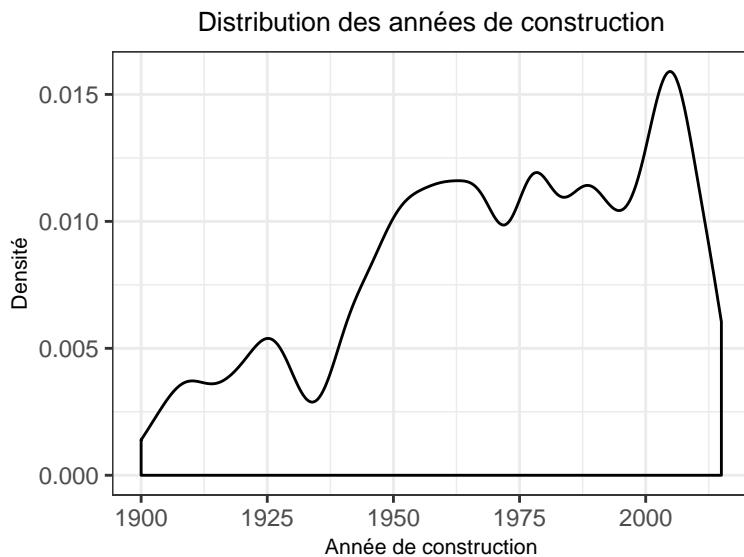
Maintenant, analysons de plus près certaines de ces variables.

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



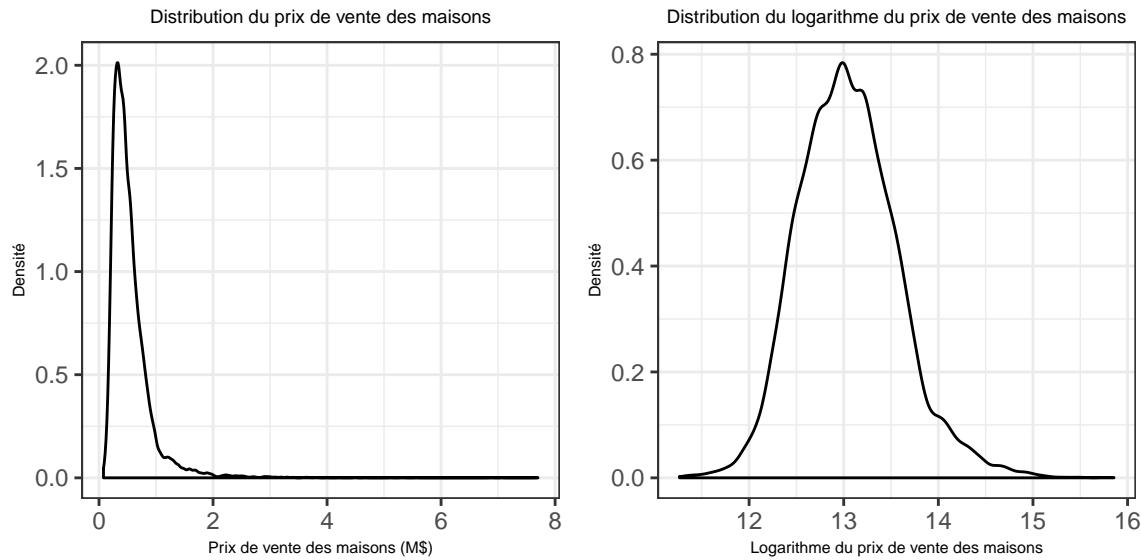
Il est possible de constater qu'il semble y avoir un effet de saisonnalité dans la vente de maisons. En effet, l'été semble être une période où il y a beaucoup de ventes de maisons tandis que l'hiver semble être une période où la vente de maisons est moins fréquente.

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 pour 1900. Ainsi, nous avons pris l'hypothèse que les données ont commencé à être recueillies en 1900.

Variable réponse



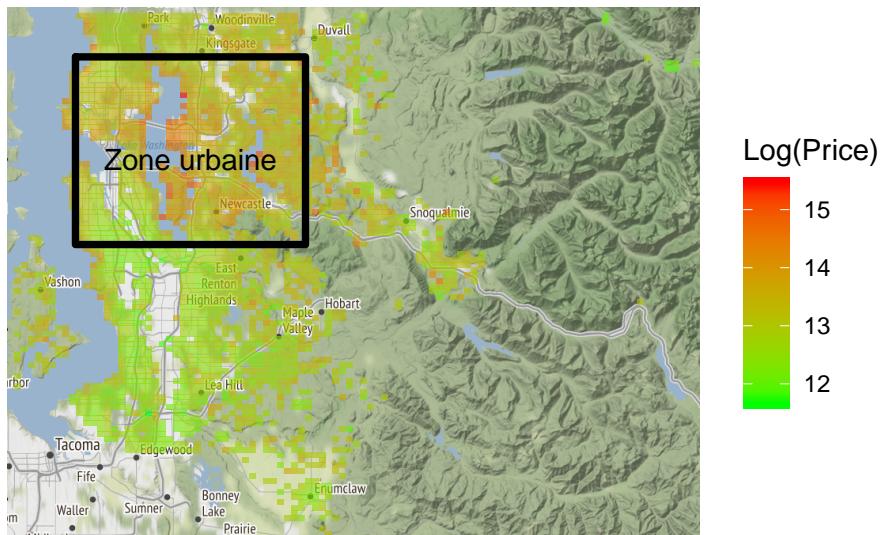
Lorsqu'on regarde le graphique de gauche, où aucune transformation logarithmique n'a été effectué sur la variable du prix, il est possible de remarquer que la distribution a une forte asymétrie à droite. Cela a pour effet de décaler la distribution à gauche de la médiane et d'étaler la queue de la distribution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

Pour pallier à 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

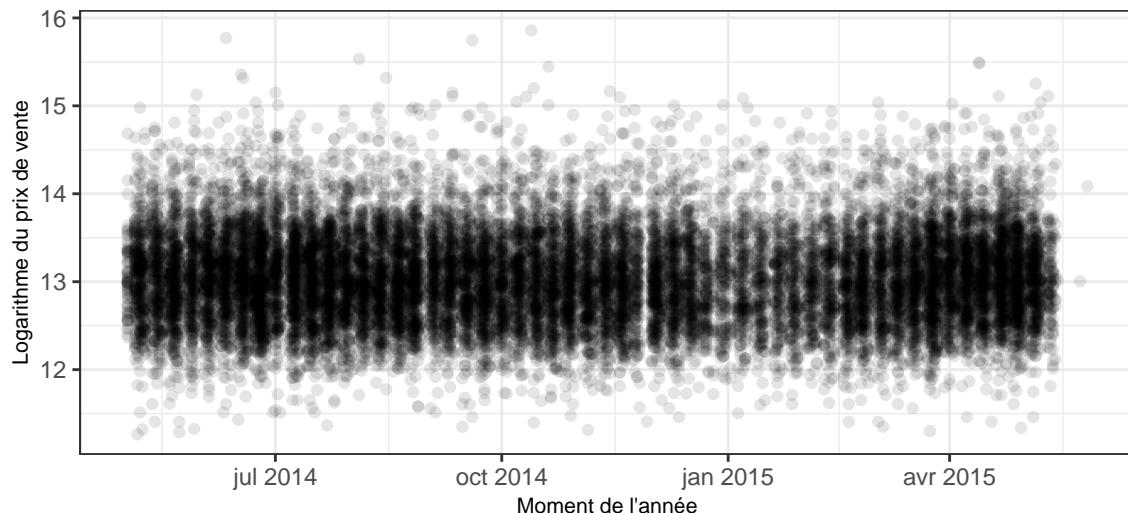
Heatmap

King County, USA



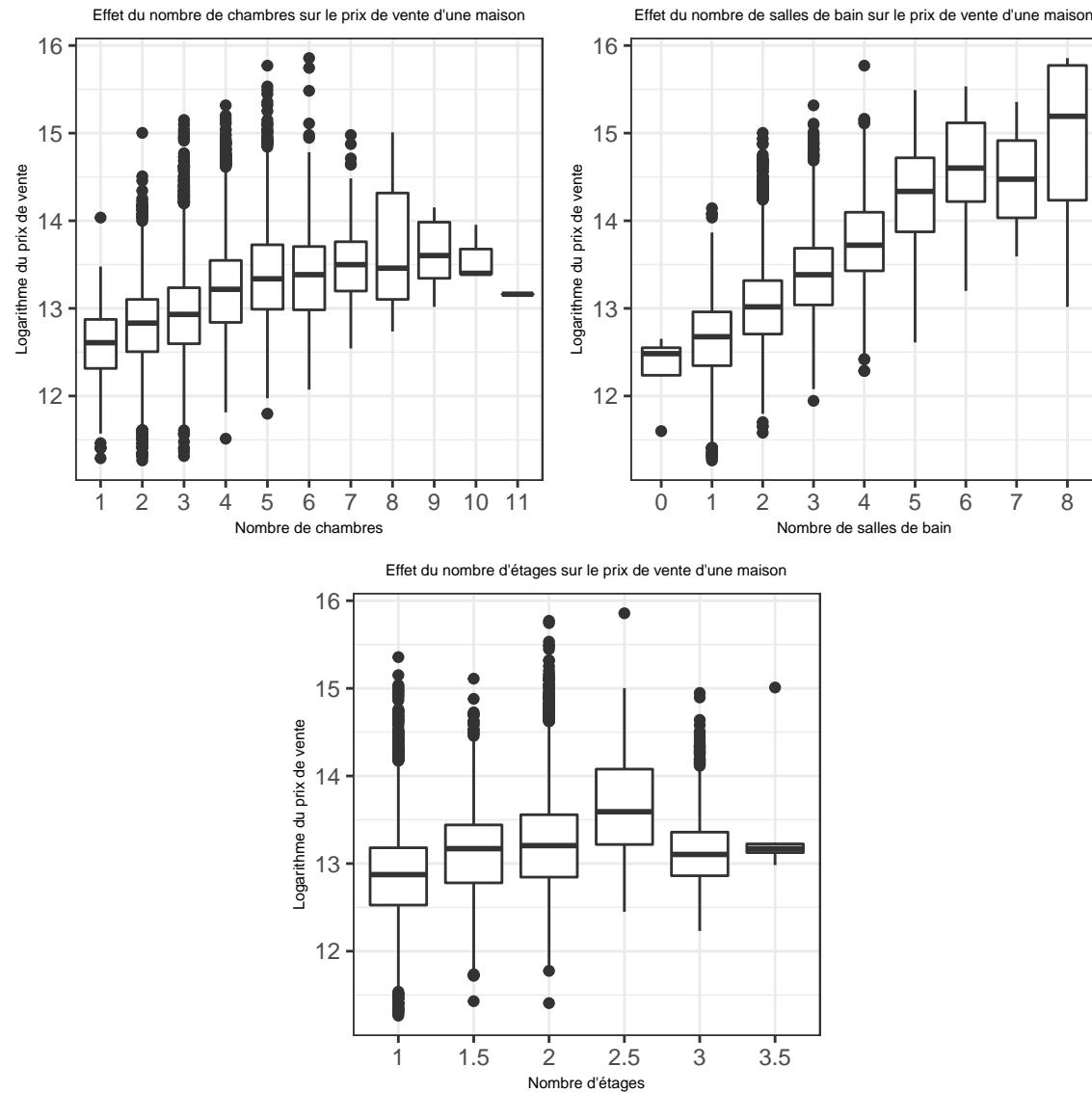
Date en fonction du logarithme du prix

Logarithme du prix de vente des maisons selon le moment de l'année



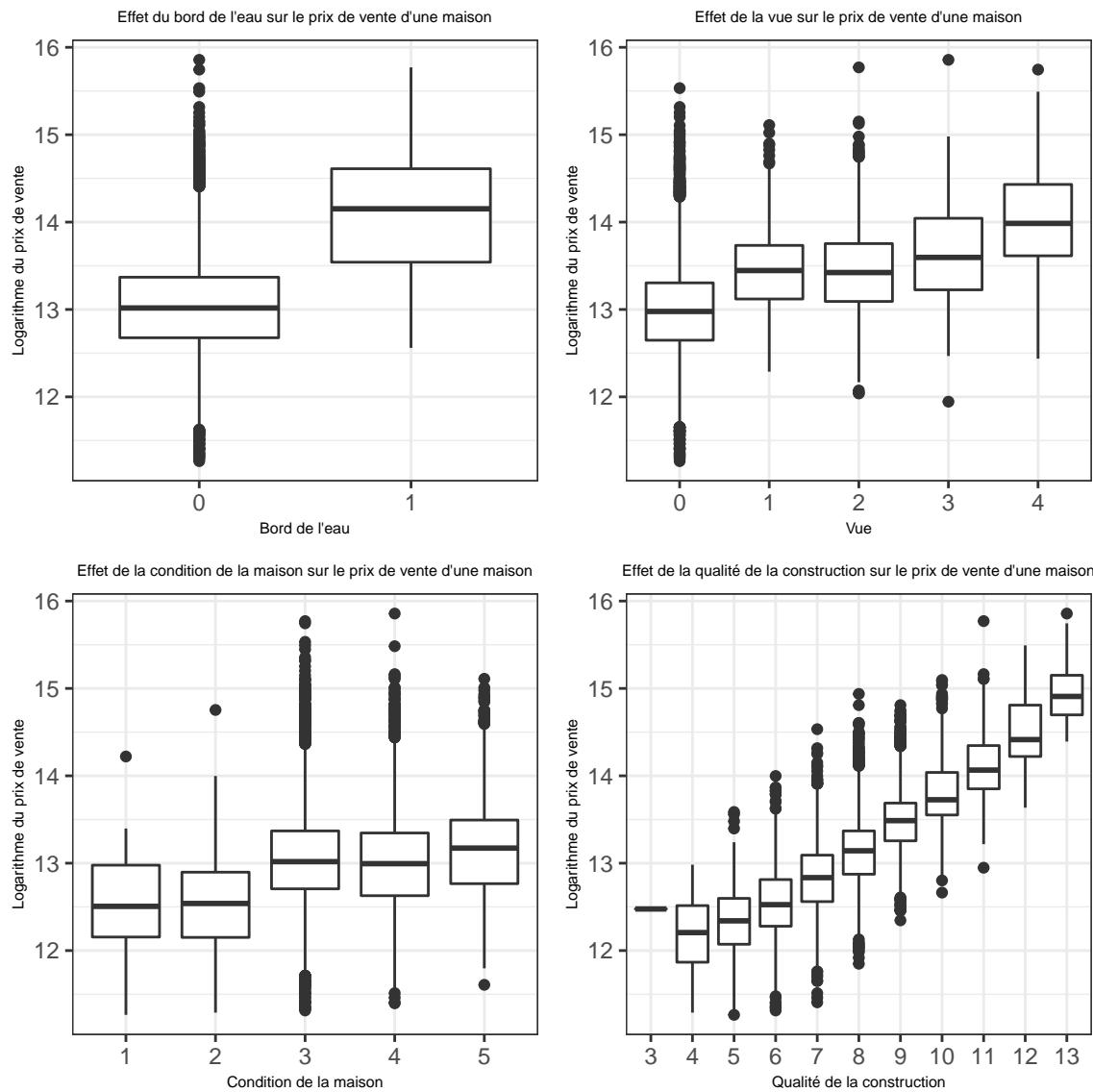
Le graphique ci-dessus montre qu'il n'y a pas de relation linéaire entre le moment de la vente de la maison et le logarithme du prix obtenu. En effet, peu importe la date, le logarithme du prix semble distribué normalement et la variance des observations est constante dans le temps. Il est intéressant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

Variables bedrooms, bathrooms et floors

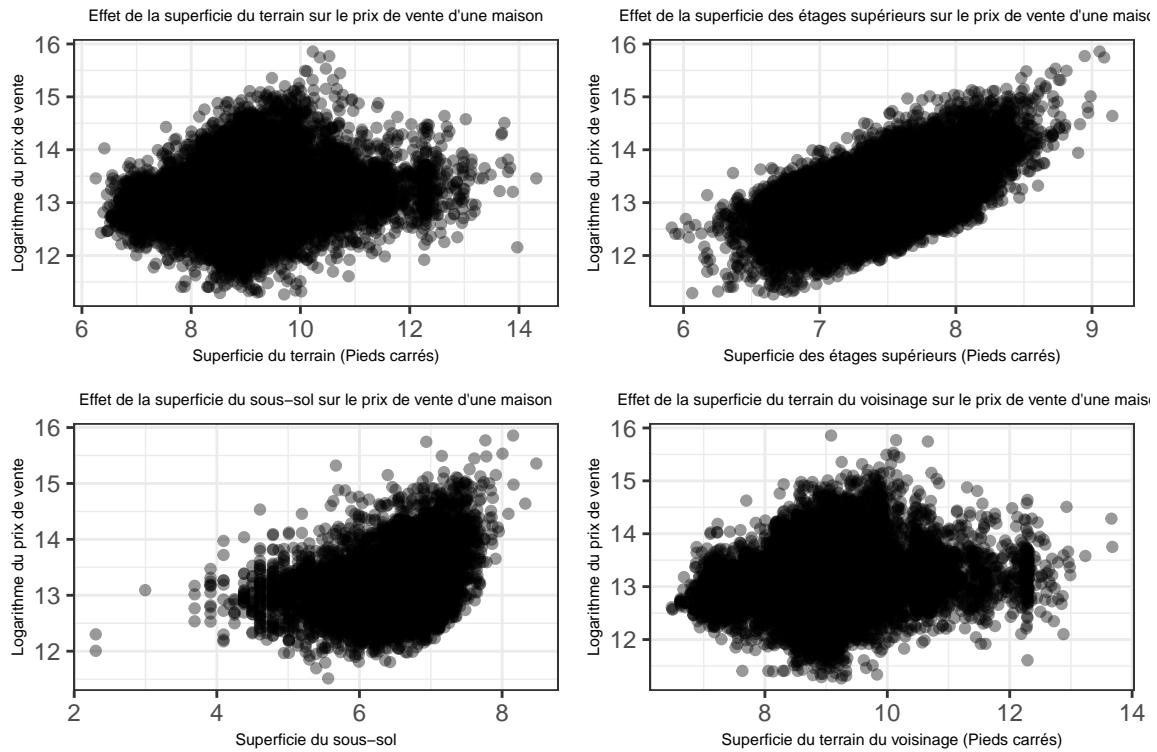


Le constat général de ces trois graphiques est que plus ces variables prennent une valeur élevée, plus le logarithme du prix augmente. Il s'agit d'un constat assez intuitif, donc ce n'est pas surprenant. Un second constat est qu'il y a une plus grande volatilité pour de faibles valeurs de ces trois variables. 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 dur 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.

Variables waterfront, view, condition et grade

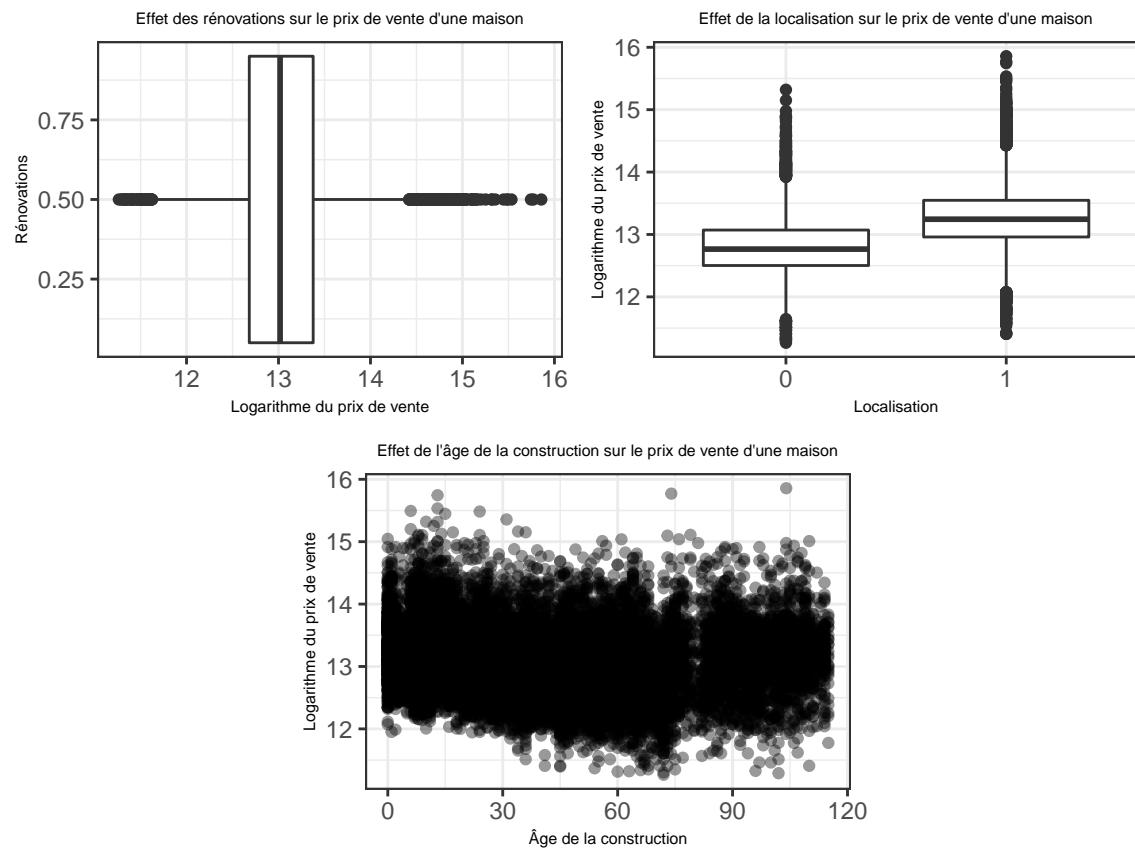


Variables de superficie



Variables age, reno et expensive area

```
## Warning: Continuous x aesthetic -- did you forget aes(group=...)?
```



Création de variables explicatives

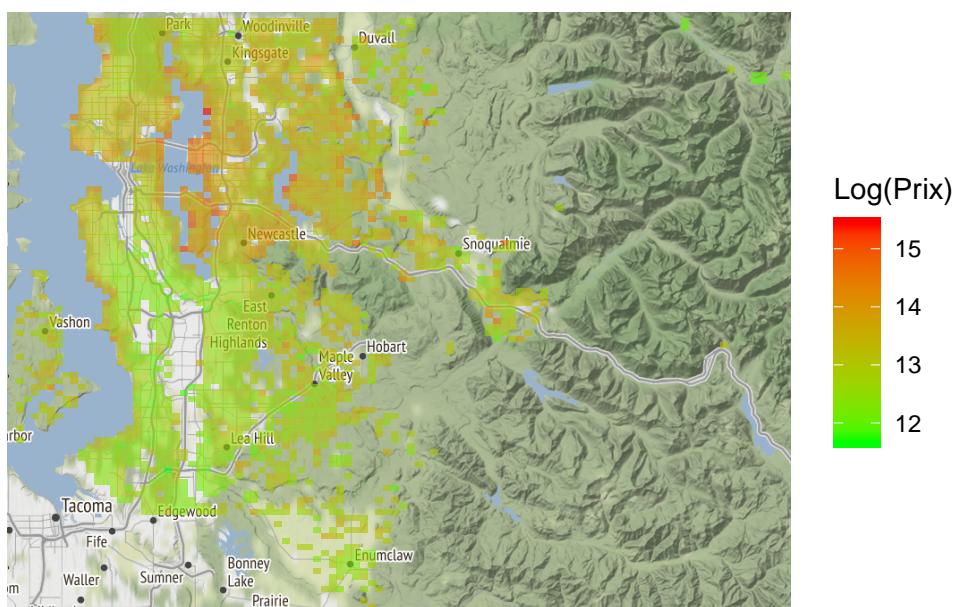
D'après les données accessibles, il a semblé pertinent d'en utiliser certaines afin de créer d'autres variables explicatives qui seront plus utiles afin de déterminer le prix de vente des maisons de King County.

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. Combiné ensemble ces deux dates crée un variable numérique discrète plus utile pour en déduire le prix d'une maison. En d'autres mots, cette variable expliquera mieux les patrons de réponse. **correct Sam ? besoin de ta validation**

La deuxième variable qui a été créée est celle à savoir si la maison a été rénové 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 similaire les uns entre les autres. Les séparer en plusieurs catégories semblent donc désuet, c'est pourquoi au final, la variable *reno* prend comme valeur 1 si la maison a déjà été rénové depuis sa construction ou 0 si elle n'a jamais été rénové.

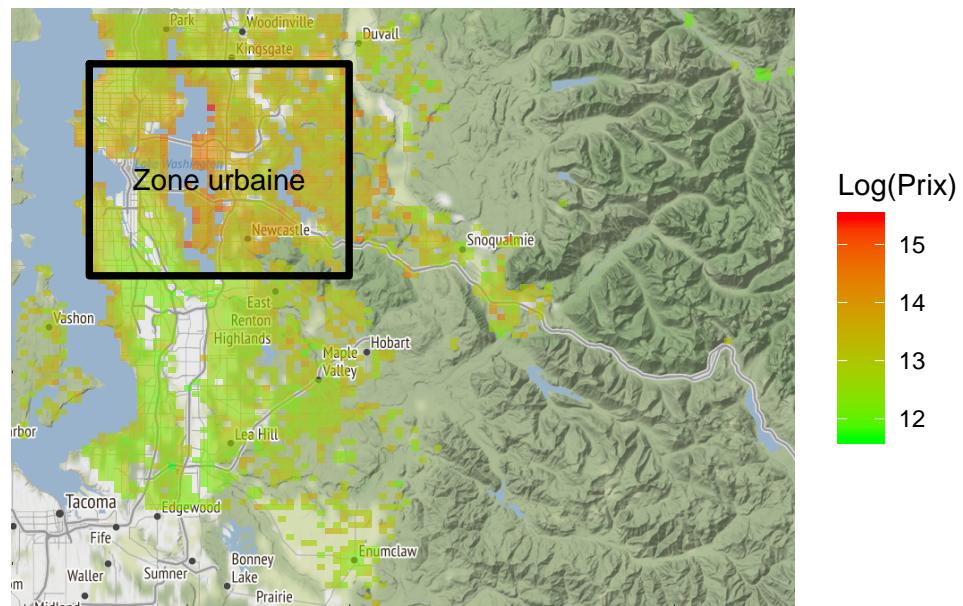
La dernière variable créée est celle représentant la région la plus coûteuse (*expensive_area*). En effet, les variables représentant la latitude et la longitude de la maison sur la carte thermique ci-dessous (**comments :faire référence ex. figure 3**) indique la position de la maison sur la planète Terre. En utilisant toutes ces données de position des maisons, il est alors facile de les situées sur une carte de la région de King County, la région à l'étude dans ce rapport. Voici d'ailleurs un aperçu des positions de chaque maison vendue à King County.

Position des maisons vendues de King County



En regardant la carte thermique affichée ci-dessus, il est possible de voir qu'une bonne partie des maisons les plus couteuses se situe au nord-ouest de la région, tout près de l'eau. Cette région semble être la zone urbaine. Il est d'ailleurs possible d'identifier la région la plus couteuse en allant chercher les coordonnées des bonnes latitudes et longitudes.

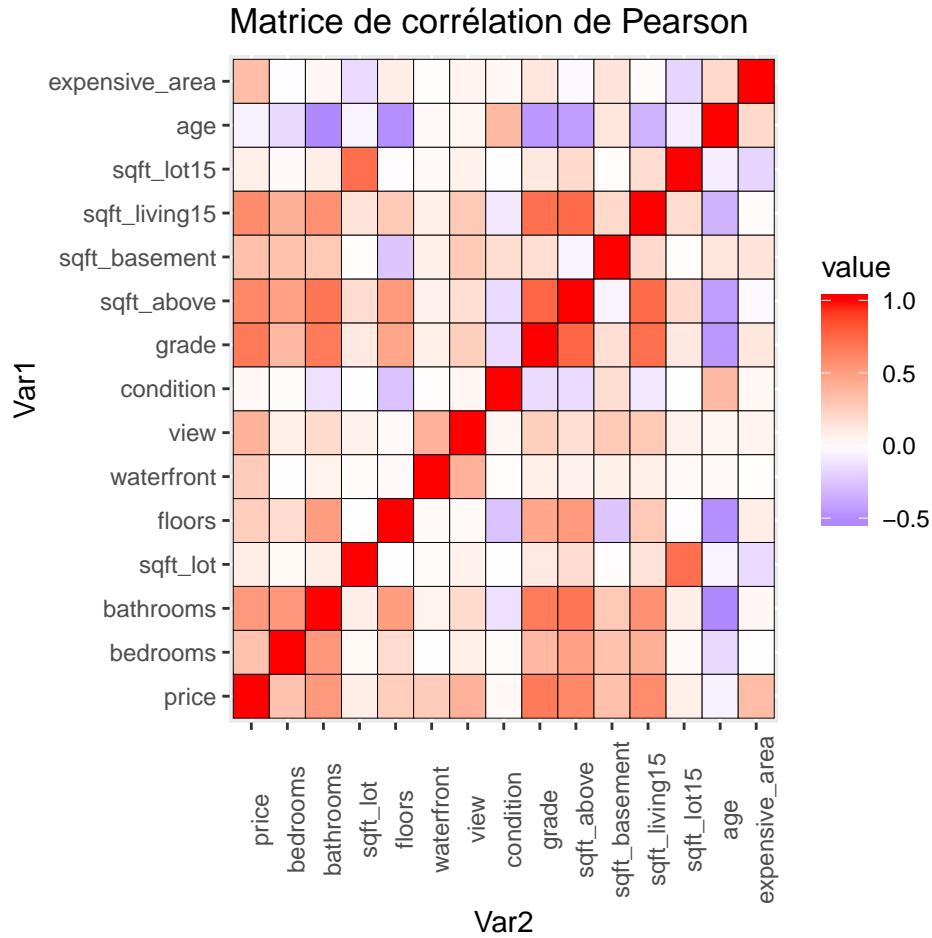
Position des maisons vendues de King County



La variable de la région la plus couteuse est une variable catégorielle qui renvoie 1 lorsque la maison est située dans la zone urbaine et renvoie 0 lorsque la maison ne se situe pas dans la zone urbaine.

Réduction de la dimensionnalité

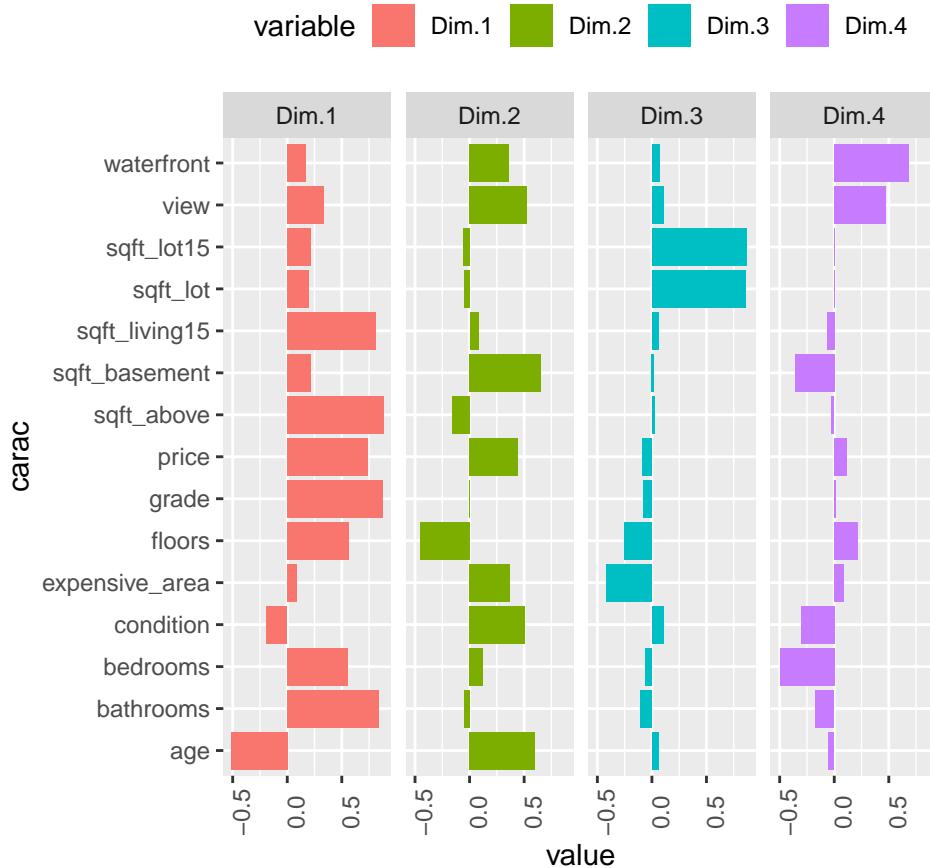
Il se trouve que malgré tout le prétraitement des données et l'analyse univariée et bivariée de celles-ci, on se retrouve avec un jeu de données contenant 16 variables différentes. Ce nombre semble élevé et le nombre de variables fait en sorte qu'on ne peut visualiser efficacement ces données en grande 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 covariances entre chaque variable du jeu de données. Encore là, il y a beaucoup de couleurs rouge et bleu qui apparaissent, signe d'une bonne corrélation entre les variables. Cependant, il est difficile d'indiquer quelle variable peut correctement compenser pour une autre. C'est pourquoi, l'ACP sera appliquée au jeu de données afin de bien résumer l'information contenue dans les 16 variables efficacement. 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 $m < n$, à des fins de réduire les besoins en ressources de calcul. Avec cette méthode, il sera possible de faire de la visualisation, puisque qu'il est facile de représenter \mathbb{R}^2 . L'ingéniosité de l'ACP consiste à maximiser la variance de la projection, limitant la perte d'information inhérente à un espace de plus faible dimension.

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. Procédons donc :

Contributions des différentes variables aux 4 premières composantes



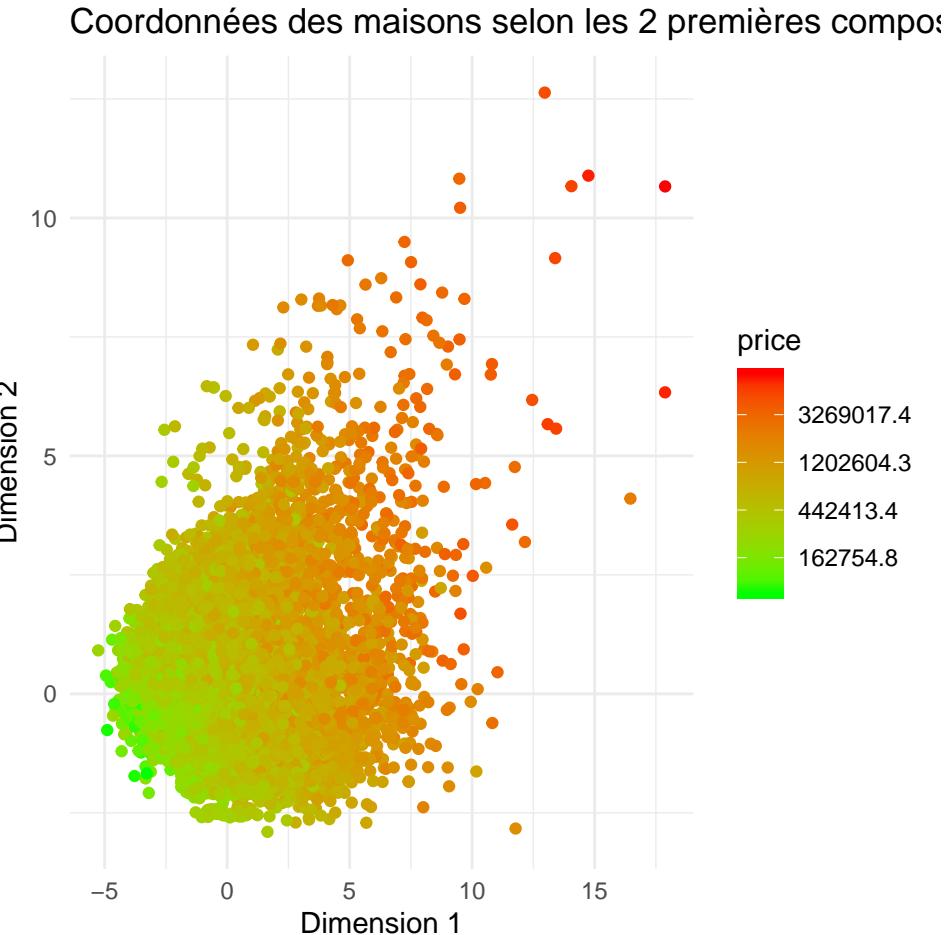
Sur le précédent graphique, il est possible de voir la contribution des variables aux 4 premières composantes créées. Commençons par analyser les deux premières composantes.

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 assumé que pour une valeur élevée de la composante 1, celle-ci prendra une valeur minime 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 et donc une petite maison ayant été construite depuis longtemps puisque la variable *age* contribue beaucoup dans les négatifs et les autres valeurs mentionnées on quant à eux plus d'importance dans le positif.

La 2e composante quant à elle semble plutôt indiquer la présence de vieilles maisons 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/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 une grosse 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 coordonnées des chacune des maisons de notre jeu de données en fonction des 2 premières composantes :



En effet, le long de l'axe des X, 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 sont associées à de grandes valeurs de cette composante et les petites maisons vieilles à des petites valeurs, sans argumenter il est donc logique de voir la différence de prix comme décrit précédemment.

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 compensées 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.

Conclusion

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)