

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
Variable réponse	3
Variable date de vente	4
Analyse bivariée	4
Heatmap	5
Date en fonction du log du prix	5
Variables	6
Variables	7
Variables	8
Variables	9
Création de variables explicatives	10
Réduction de la dimensionnalité	12
Conclusion	13
Bibliographie	14
Annexe	15

Introduction

Dans le cadre du travail, le prix de ventes des maisons dans la région de Seattle (King County) sera modélisé en utilisant de nombreuses caractéristiques ayant une incidence sur la valeur d'une maison. Le prix de ventes 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 un estimé des montants de prêts hypothécaires dans une région donné en se basant sur le prix de vente des maisons pour mieux gérer le risque de la compagnie. La pertinence de trouver cette variable qu'est le prix de ventes 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 une informations pertinentes sur les variables exogènes ainsi que sur la variable réponse sous forme d'une analyse univariée et bivariée.

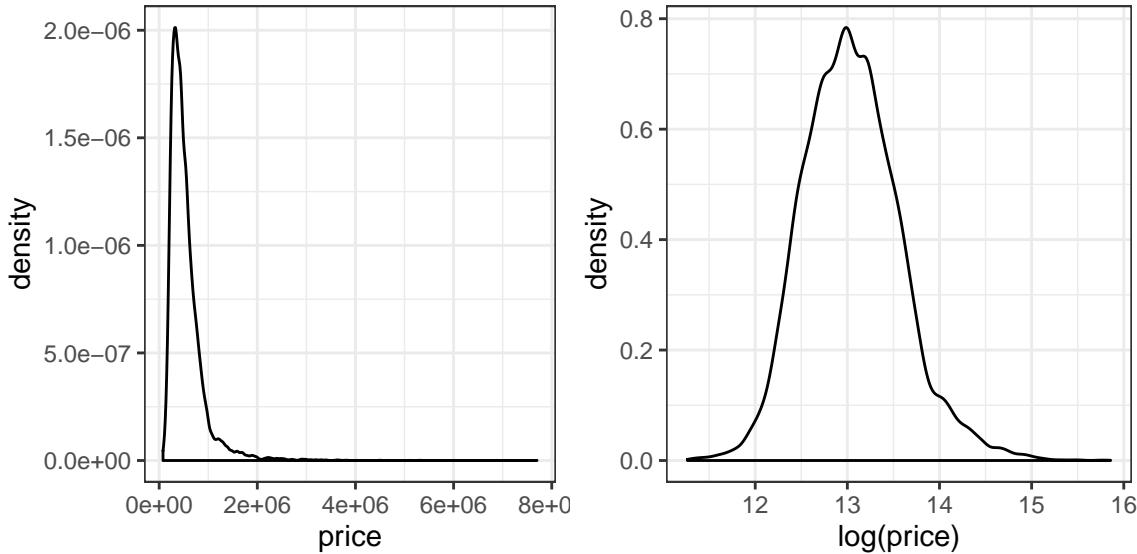
Traitement des erreurs

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 remplis, donc il ne s'agit pas de données manquantes. De plus, comme ces données contenaient toutes un espace de terrains et qu'elles représentaient une faible proportion, 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. C'est pourquoi le nombre de chambres pour cette observation a été mis à 3.

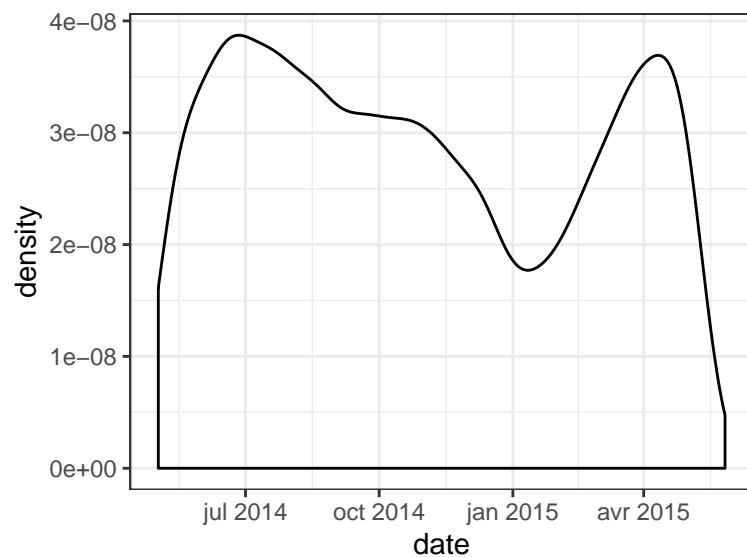
— Année de maison à 1900 (à traiter dans l'analyse univarié)

Analyse univariée

Variable réponse

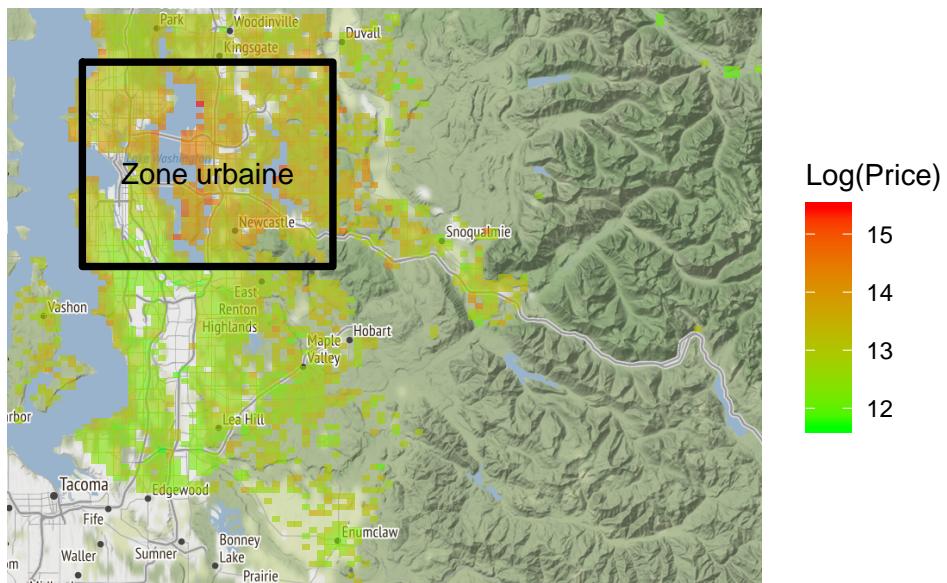


Variable date de vente

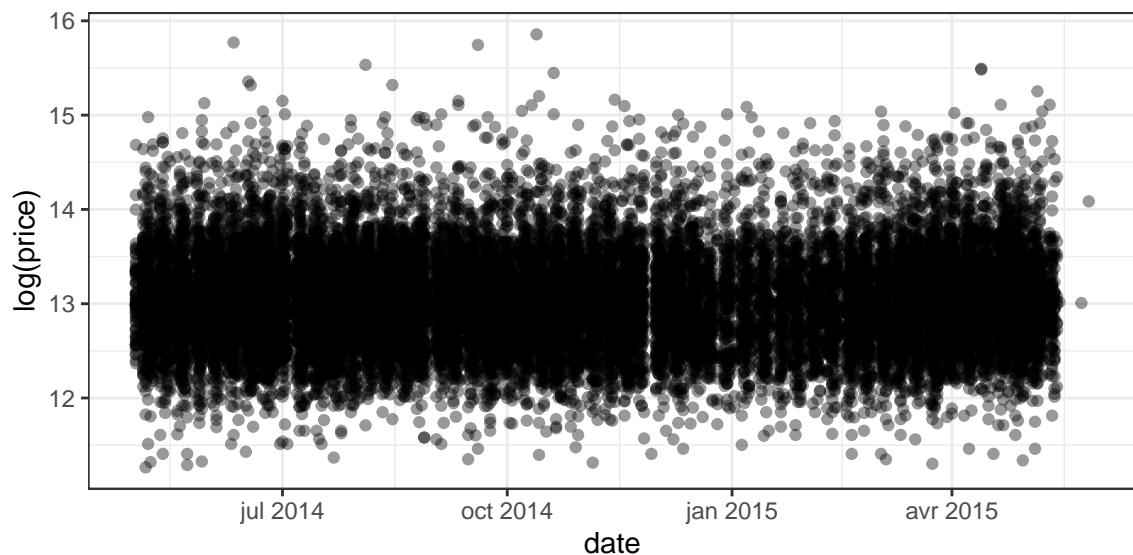


Analyse bivariée

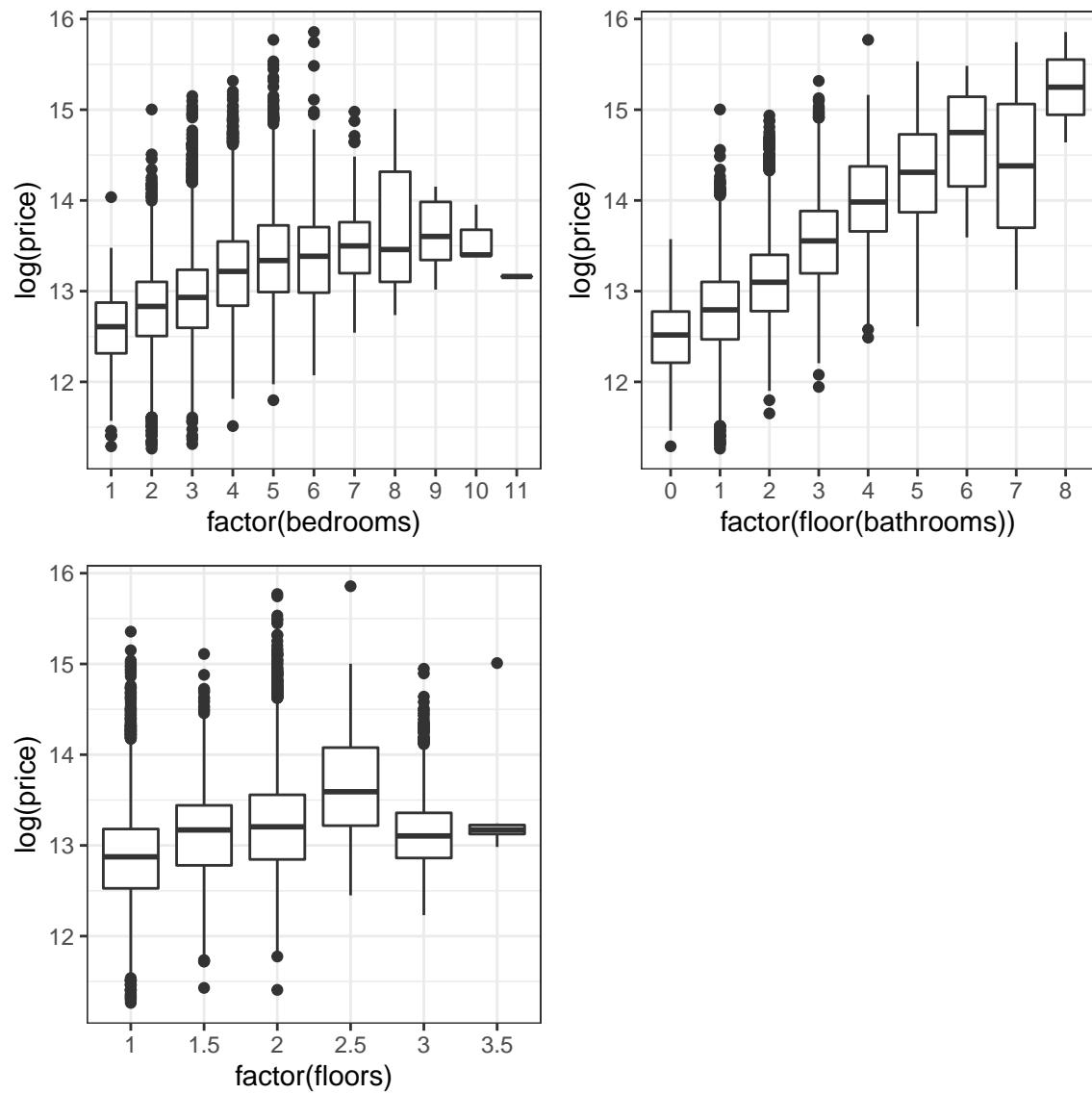
Heatmap



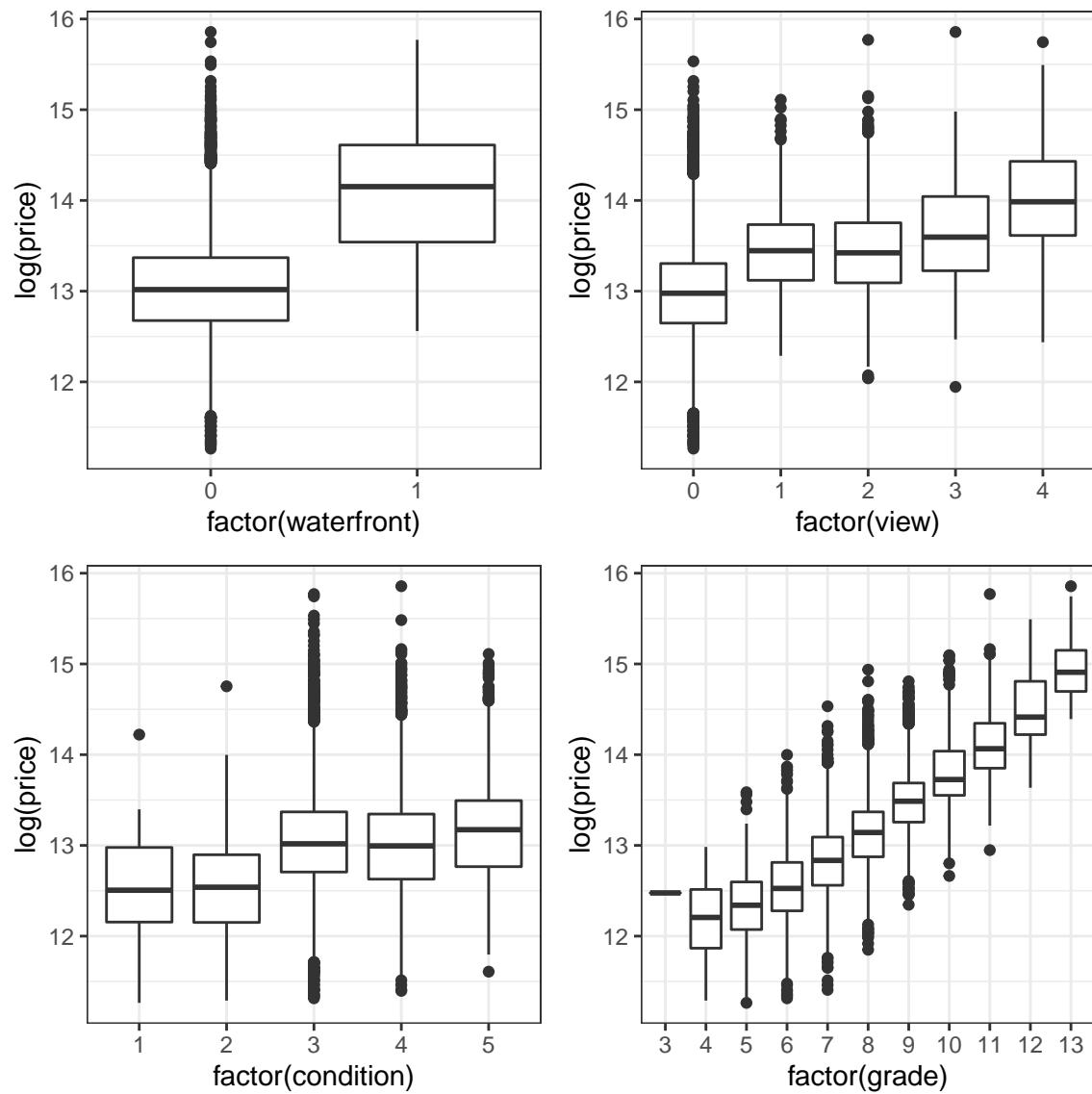
Date en fonction du log du prix



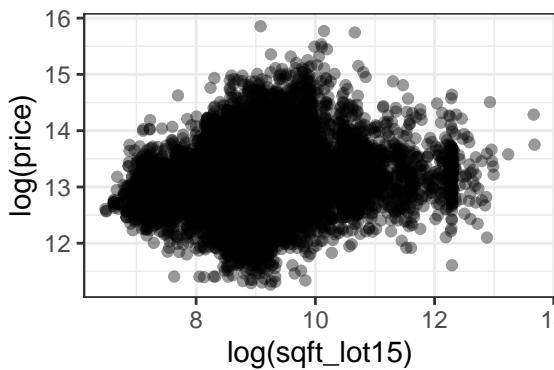
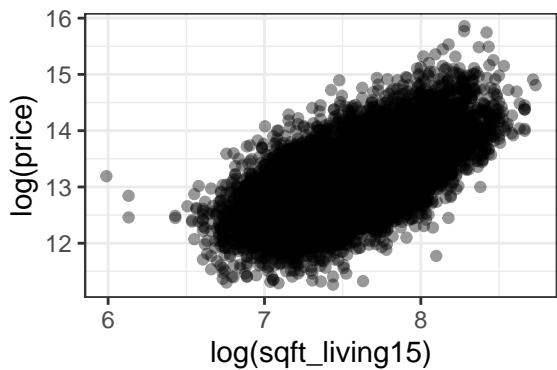
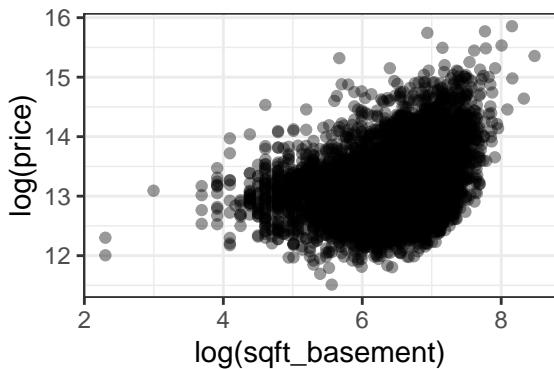
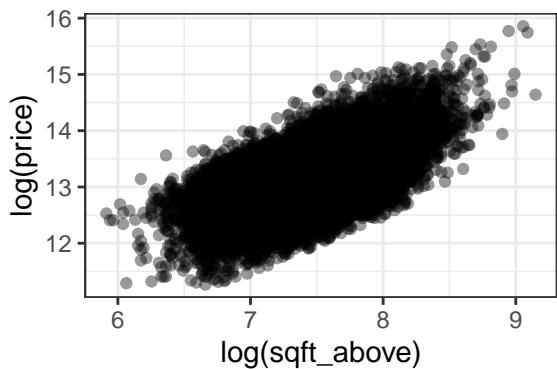
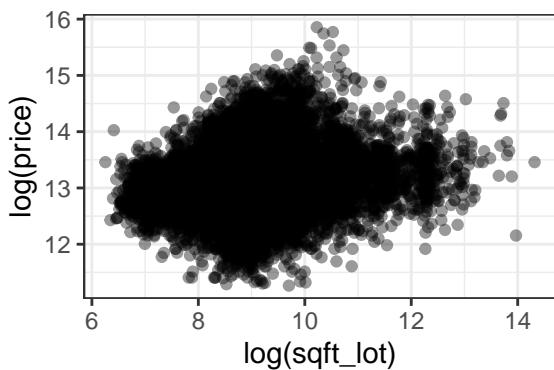
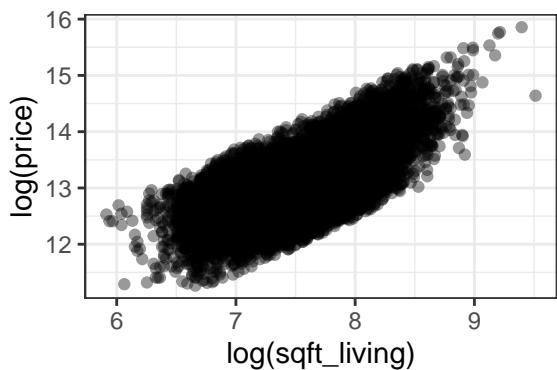
Variables ..



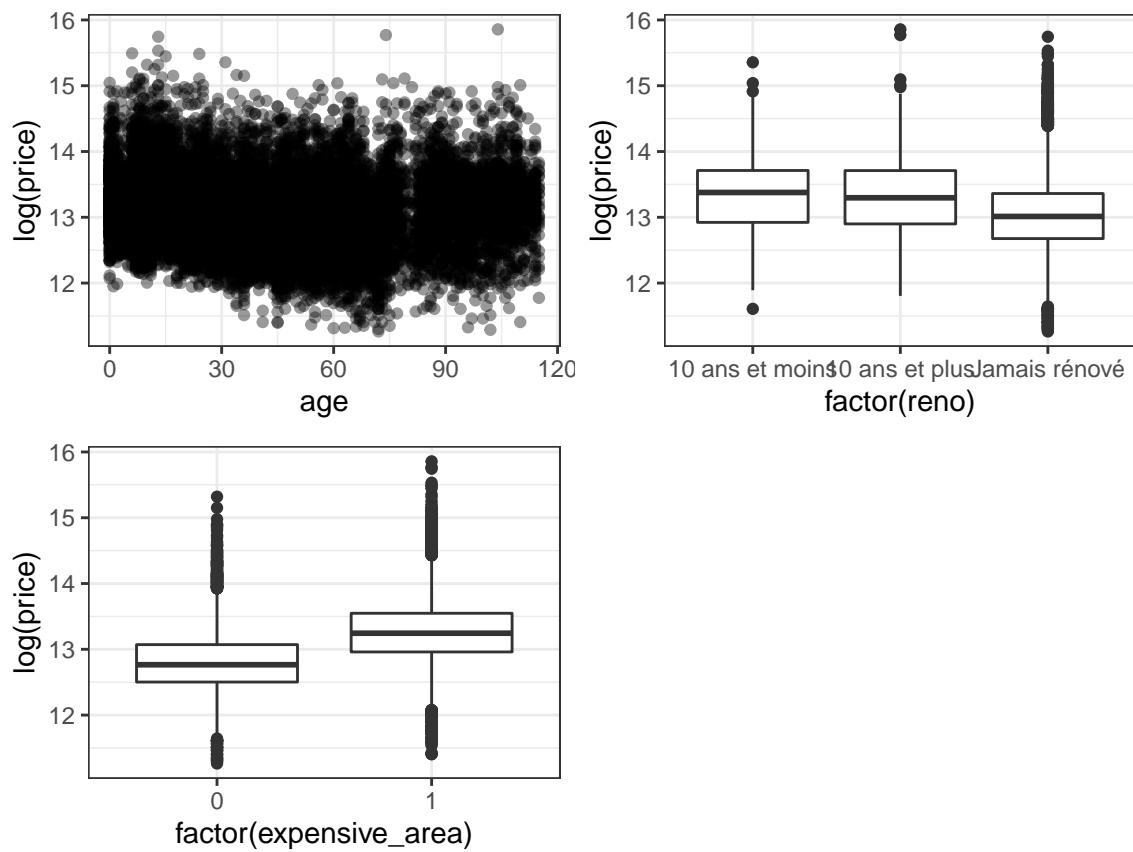
Variables ..



Variables ..



Variables ..



Création de variables explicatives

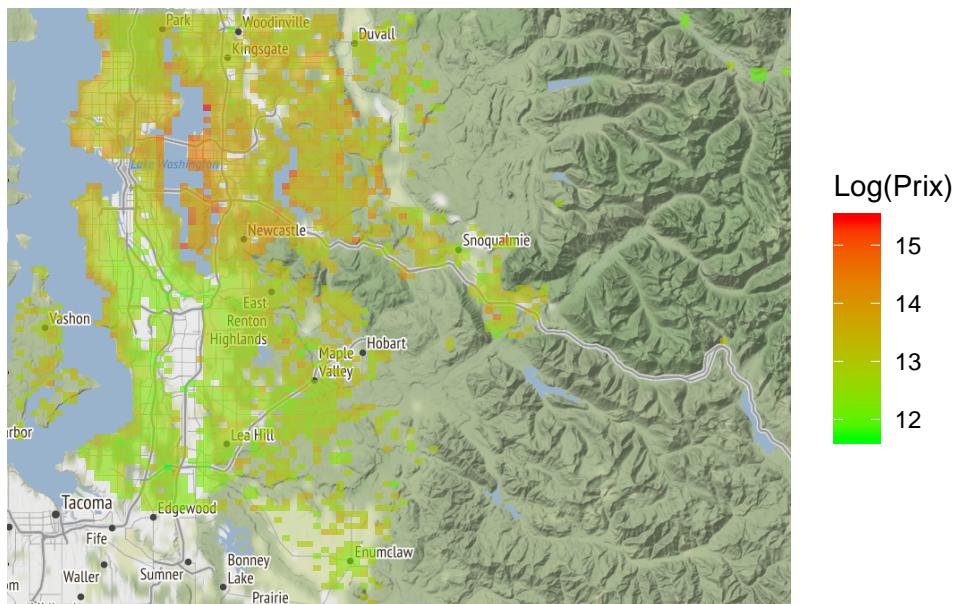
D'après les données accessibles, il a semblé pertinent d'en utiliser certaines afin de créer d'autres variables explicatives qui seront plus utiles afin de déterminer le prix de vente des maisons de King County.

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. **Comments :On pourrait dire que c'est plus facile analyser l'age vs yr_built, que c'est plus représentatif. Ca s'expliquerait mieux à des patrons. (T'es pas obligé d'ajouter ça si t'aimes pas !)**

La deuxième variable qui a été créée est celle du nombre d'années depuis la dernière rénovation de la maison (*reno*). Cette variable a d'ailleurs été calculée de la même manière que l'âge de la maison décrite précédemment. La date de vente de la maison de la variable *date* et l'année de rénovation de la maison de la variable *yr_renovated* ont été utilisés. Par contre, étant donné le nombre élevé de maisons n'ayant jamais été rénové dans le jeu de données initiales (20699 comparativement à 914), il a été décidé d'en faire une variable catégorielle ordinale. Il s'agit d'une variable qui se divise en 3 sous-groupes, le premier étant *10 ans et moins* pour les maisons ayant eu une rénovation dans les 10 années précédant leur vente, le deuxième étant *10 ans et plus* pour les maisons ayant eu une rénovation il y a 10 années ou plus et le dernier étant *Jamais rénové* pour les maisons n'ayant jamais été rénovées depuis leur construction.

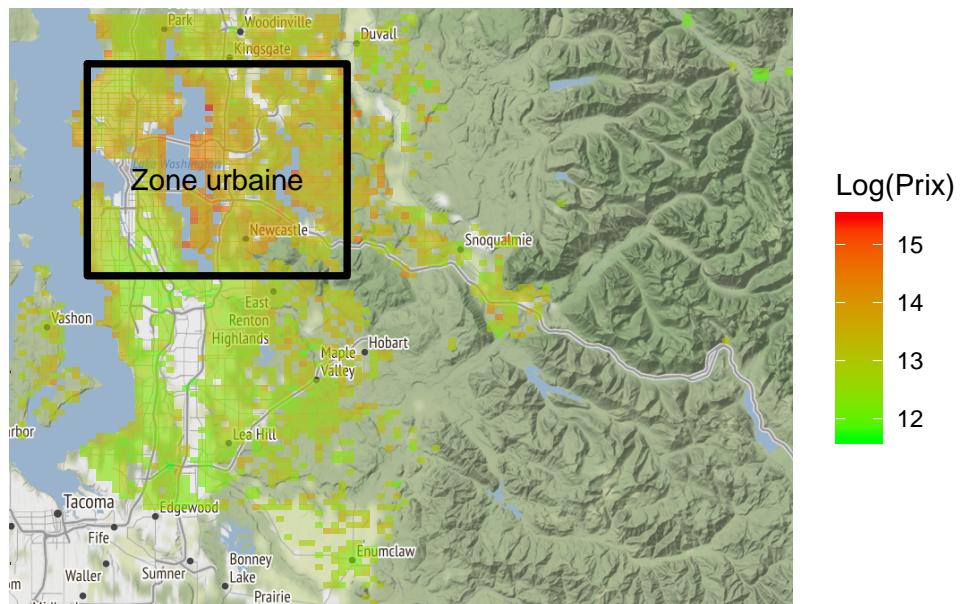
La dernière variable créée est celle représentant la région la plus coûteuse (*expensive_area*). En effet, les variables représentant la latitude et la longitude de la maison sur la carte thermique ci-dessous (**comments : faire référence ex. figure 3**) indique la position de la maison sur la planète 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 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

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)