

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
Autres sections ?	10
Conclusion	11
Bibliographie	12
Annexe	13

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 utilisé par une compagnie d'assurance pour des enjeux en assurance habitation ou en gestion des risques (prêts hypothécaires). 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.

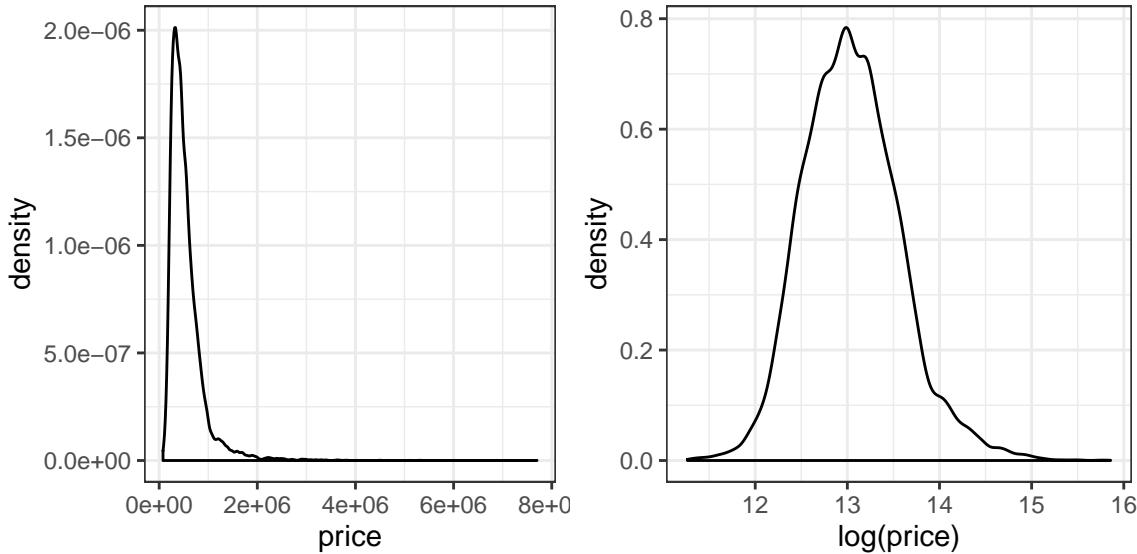
Traitement des erreurs

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.

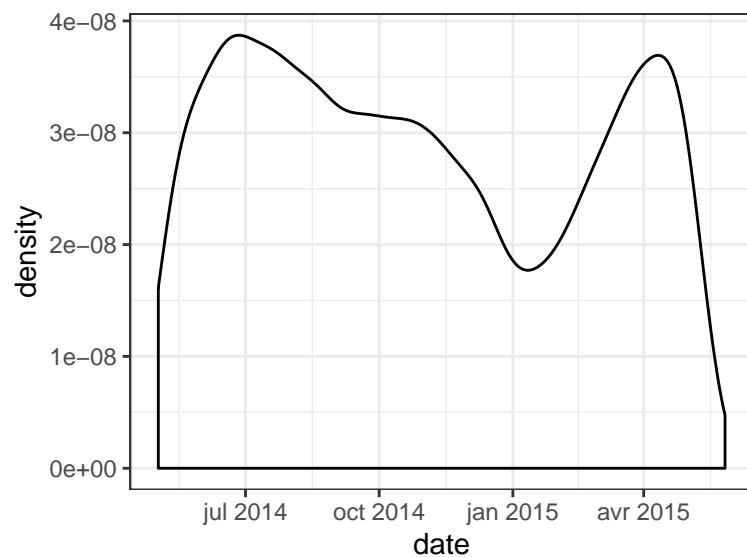
— Année de maison à 1900 (à traiter dans l'analyse univarié)

Analyse univariée

Variable réponse

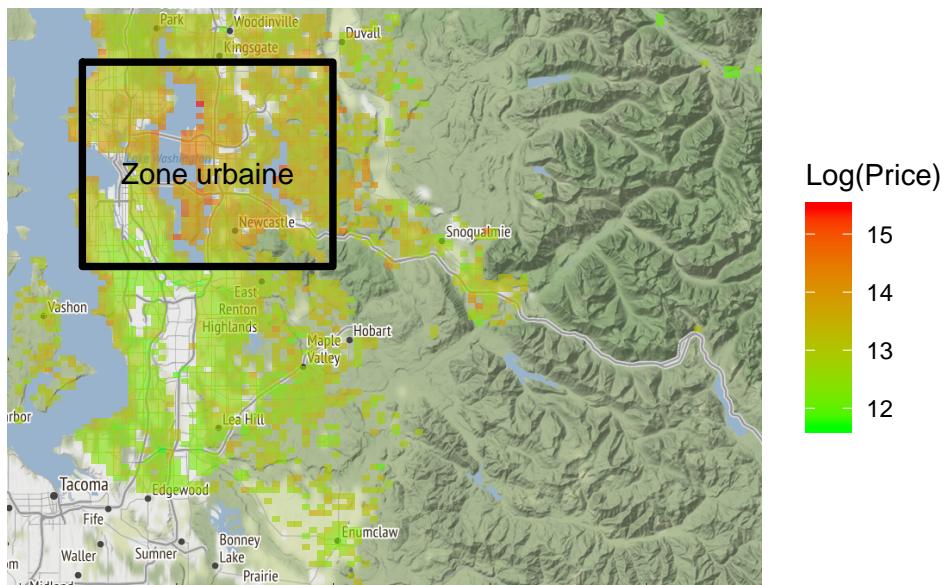


Variable date de vente

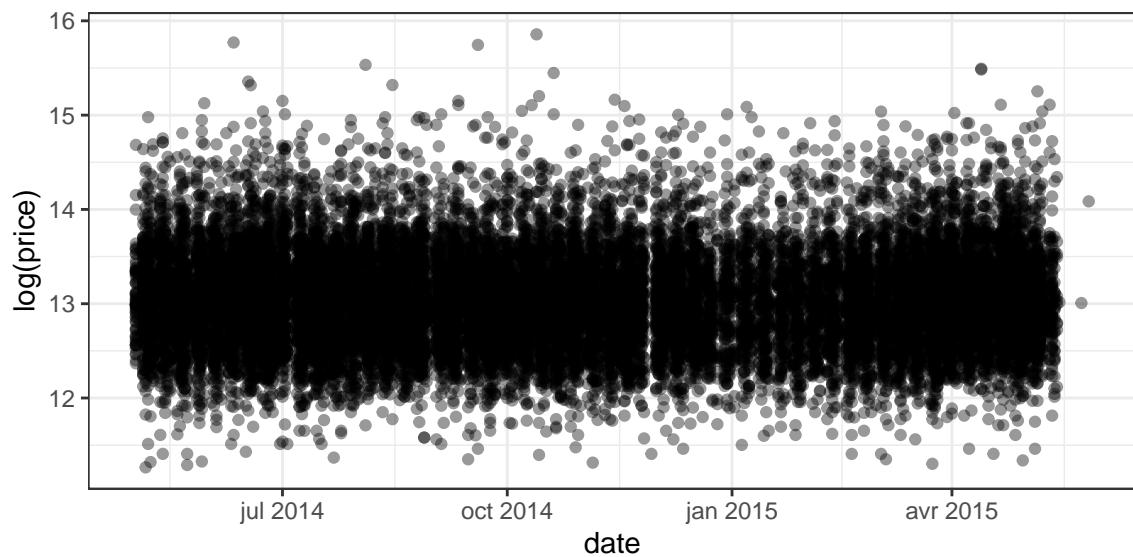


Analyse bivariée

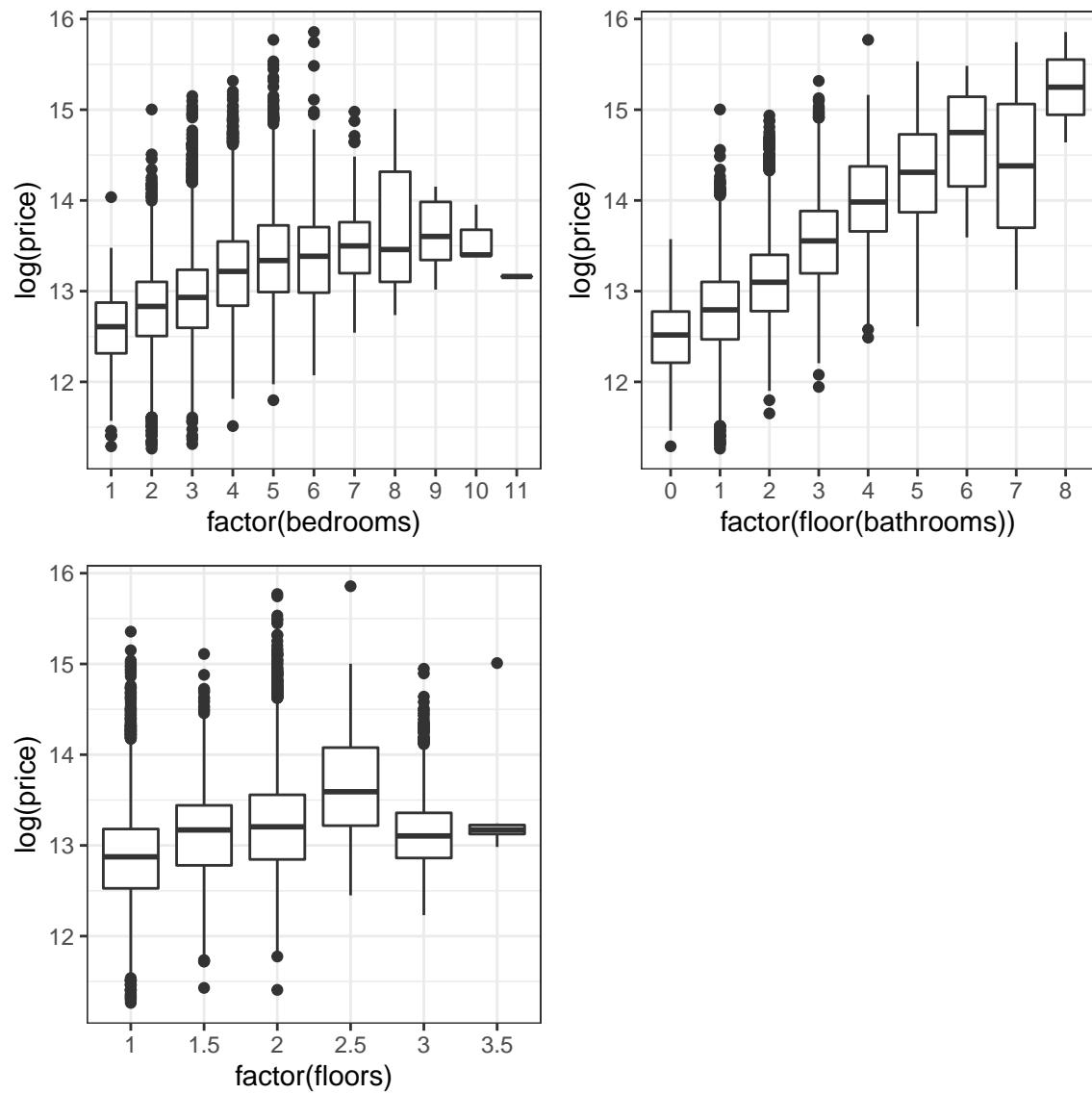
Heatmap



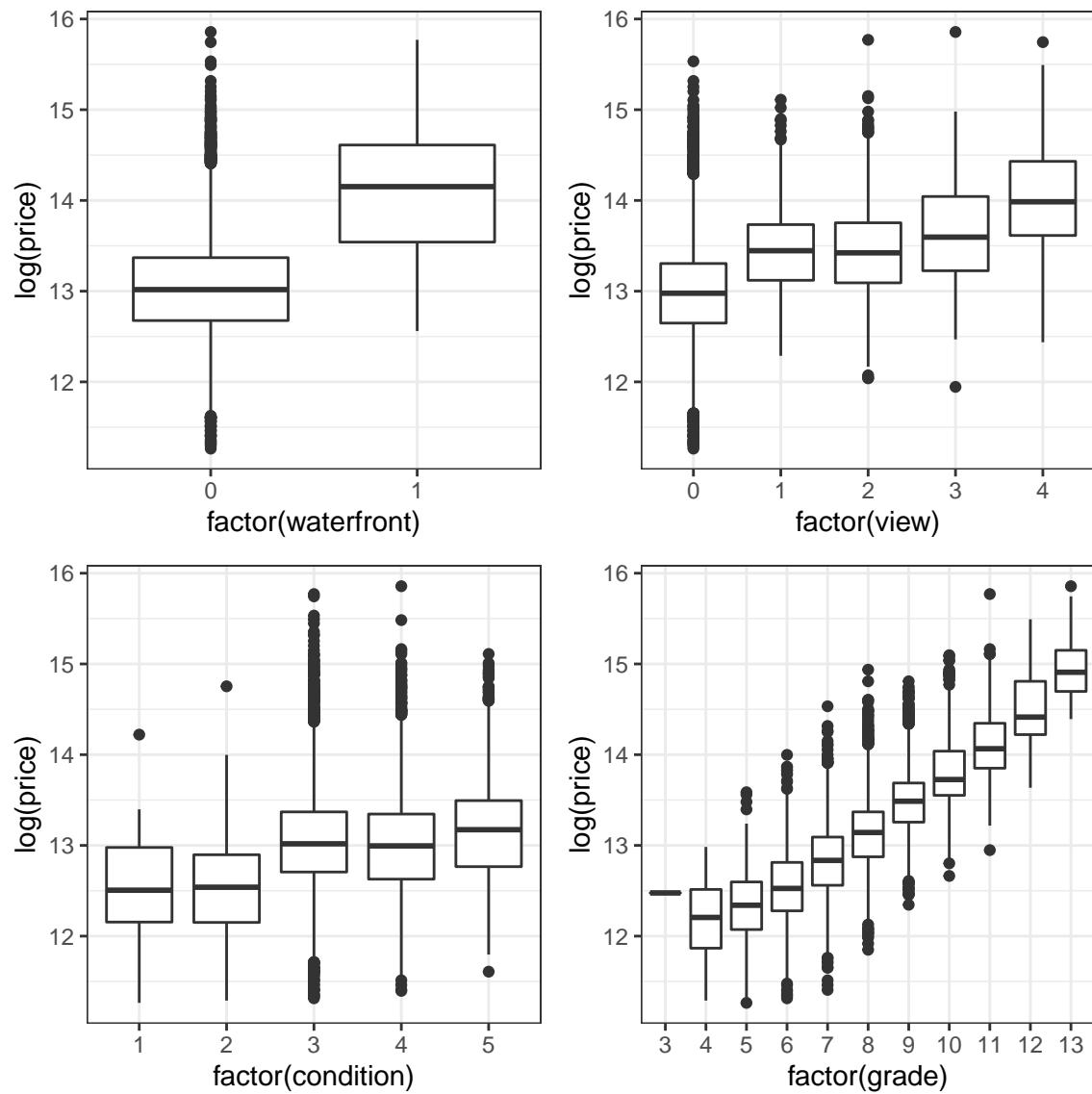
Date en fonction du log du prix



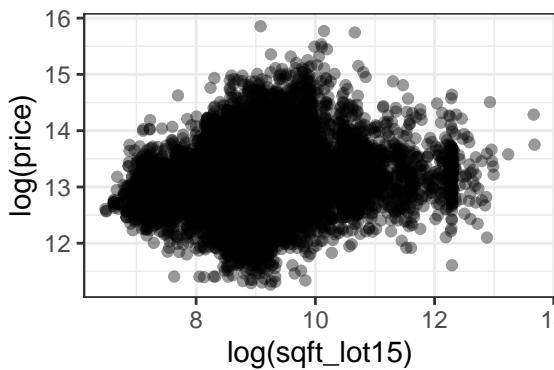
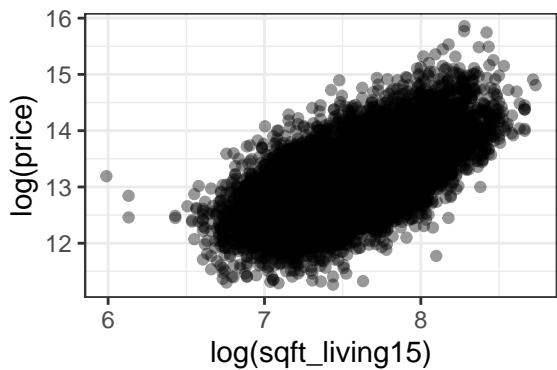
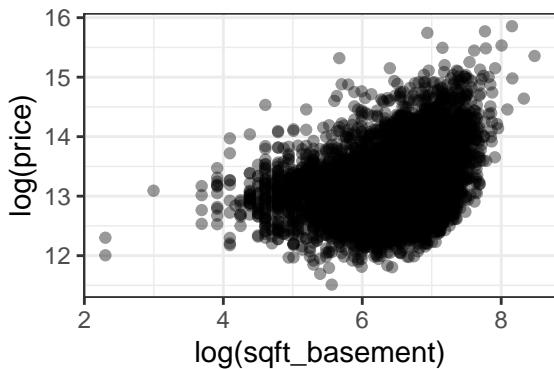
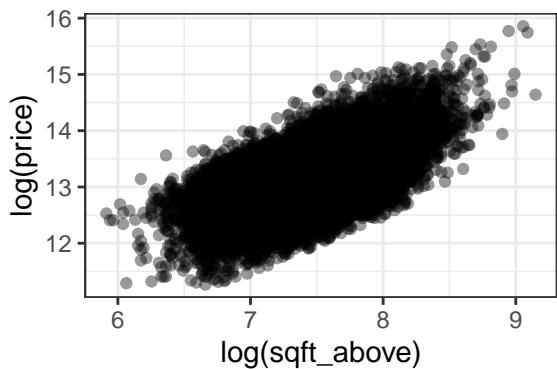
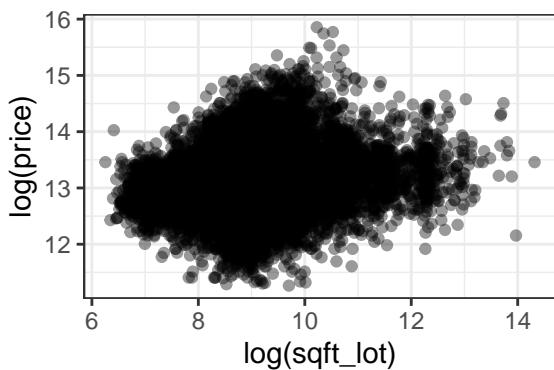
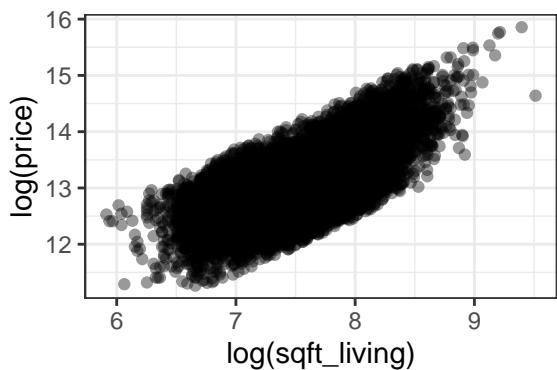
Variables ..



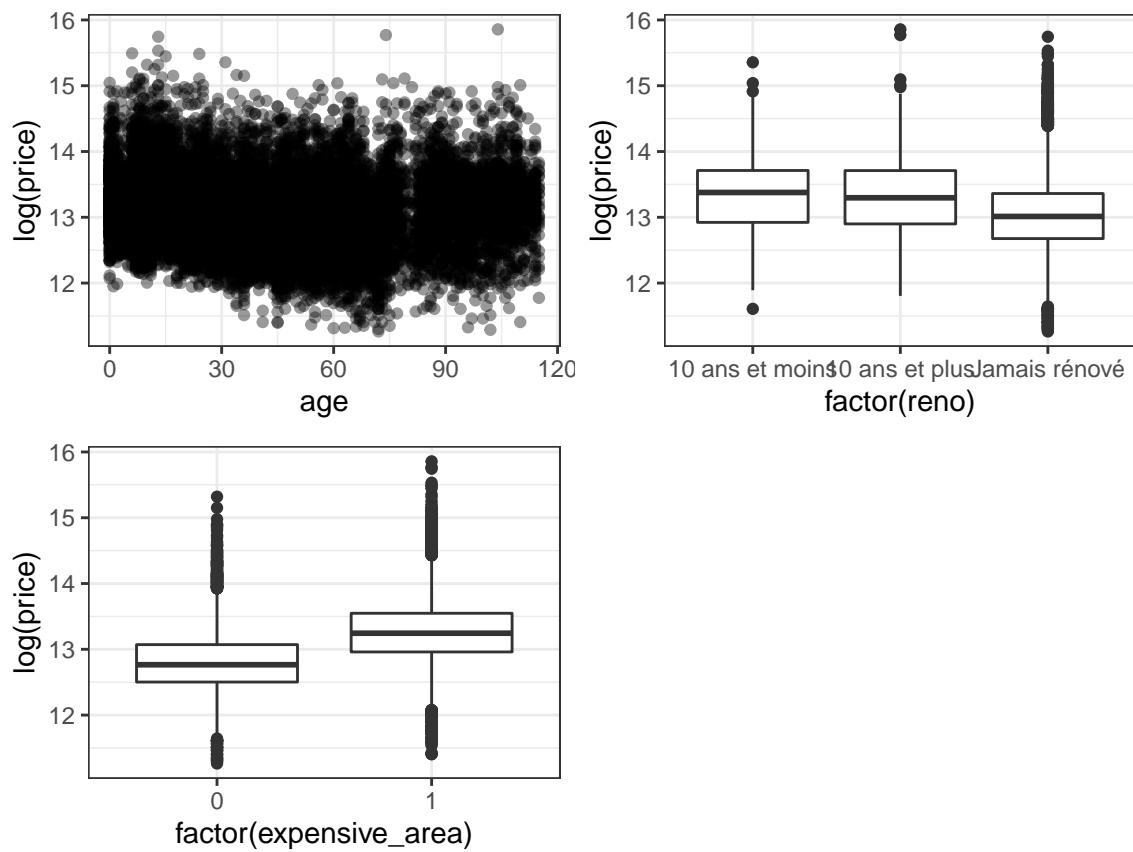
Variables ..



Variables ..



Variables ..



Autres sections ?

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)