Rapport – Analyse airbnb Seattle

Amine Halhali, Ismail Taaissat, Mohammed-Ibrahim Benahmida, Anas Imarrighen

2025-11-02

Contents

1 Introduction **2**

1.1 Contexte de l'étude 2

1.2 Objectifs 2

2 Description du jeu de données **3**

3 Analyse descriptive **4**

3.1 Statistiques de base 4

3.2 Distribution des variables clés 5

3.3 Distribution de la satisfaction 6

3.4 Distribution des types de logement 7

4 Analyse exploratoire bivariée **8**

4.1 Corrélations entre variables numériques 8

4.2 Visualisation : boite a moustache 9

5 Modélisation **10**

5.1 Régression linéaire multiple 10

5.2 Régression logarithmique 10

6 Analyse géographique **11**

6.1 Cartographie des prix par coordonnées 11

7 Interprétation et conclusions **12**

7.1 Facteurs influençant le prix 12

7.2 Limites de l'analyse et recommandations 12

1 Introduction

1.1 Contexte de l'étude

Le marché des locations de courte durée via des plateformes comme airbnb connaît un développement rapide dans les grandes villes. Comprendre les facteurs qui influencent le prix d'une location est essentiel pour les propriétaires souhaitant optimiser leurs revenus, mais aussi pour les voyageurs souhaitant comparer les offres.

1.2 Objectifs

L'objectif de cette analyse est donc de pouvoir observé le lien de corrélation et/ou de causalité entres les paramétres mis à disposition dans le jeu de donnée. Nous nous attarderons sur les points important comme le prix et les avis comme base pour notre étude afin de bien répondre à la problématique: Quels facteurs influencent le prix d'une location airbnb ?

Table 1: Visualisation des 20 premières lignes du fichier csv

price	reviews	overall_satisfaction	bedrooms	bathrooms	accommodates
250	21	5.0	4	2.5	8
100	1	NA	2	1.0	4
82	63	4.5	1	1.0	2
49	462	5.0	0	1.0	2
90	134	4.5	1	1.0	2
65	130	4.5	1	3.0	2
78	401	5.0	1	1.0	2
165	35	5.0	2	1.0	4
95	36	5.0	2	1.0	3
115	76	4.5	1	1.0	4
135	57	5.0	2	1.0	4
50	36	4.5	0	1.5	1
41	94	4.5	1	NA	2
109	20	5.0	1	1.0	4
250	80	5.0	3	2.0	8
37	125	4.5	1	1.5	2
50	70	4.5	3	1.5	6
100	17	4.5	1	1.0	1
72	89	5.0	0	1.0	2
60	78	4.5	1	1.0	2

2 Description du jeu de données

On compte 17 paramétres différent dans le jeu de données:

- room_id : correspond à la clé primaire de la table (l'identifiant unique du n-uplet)
- host_id : correspond à une clé étrangère surement celle d'une table propriétaire
- room_type : le type de chambre on en compte 3 différentes(Chambre privée,chambre partagé,appartement)
- adress : l'adresse de la ville dans laquelle le airbnb se trouve
- reviews : le nombre d'avis sur le airbnb
- overall_satisfaction : la note moyenne du airbnb
- accommodates : capacité de logement en nombre de personnes
- bedrooms : le nombre de chambre
- bathrooms : le nombre de salle de bains / toilettes (certains sont à virgule et désigne des salles de bains et toilettes séparés)
- price : le prix de la location en lien avec la devise et le temps de réservation
- last_modified : la dernière modification du airbnb
- latitude : cet élément représente la latitude du airbnb
- longitude : cet élément représente la longitude du airbnb
- location : correspond a la version de latitude et de longitude reconnaissable par la base de donnée avec une table PostGIS
- name : le nom du bien immobilier sur airbnb
- currency : la devise utilisé pour le paiement
- rate_type : le mode de facturation

3 Analyse descriptive

3.1 Statistiques de base

Table 2: Statistiques descriptives des variables numériques

Variable	N	Mean	SD	Median	Min	Max
price	7576	113.02	122.48	88	15	5900
reviews	7576	47.66	65.89	21	0	687
bedrooms	7576	1.39	1.01	1	0	8
bathrooms	7574	1.31	0.64	1	0	8
accommodates	7576	3.68	2.33	3	1	28

Sur ce tableau nous pouvons observer une multitude d'informations:

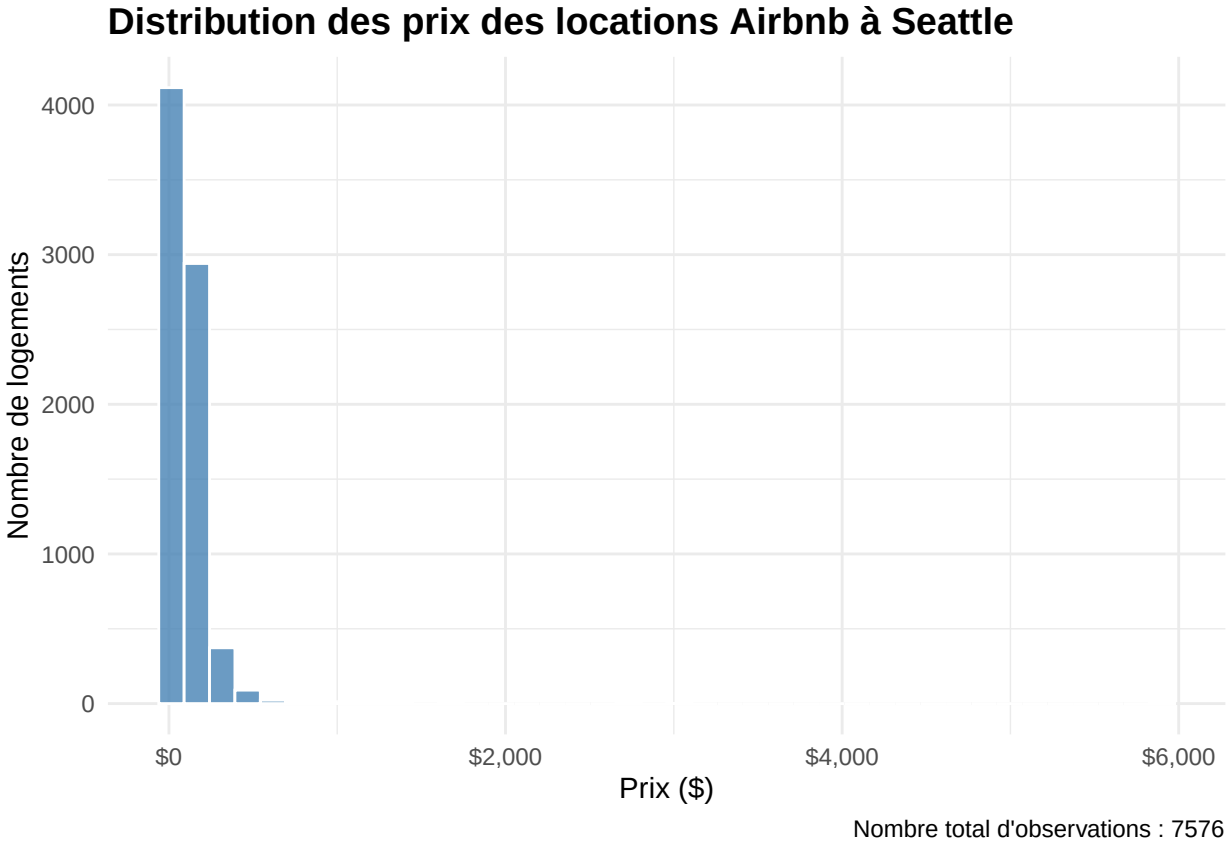
Premièrement on voit que le fichier contient 7576 n-uplet ce qui va un peu compliquer la visualisation graphique des données.

Des moyennes importantes sont aussi visible comme celle des reviews qui est assez élevé et celle du prix à la nuit qui est de 113 dollars américain.

D'autres donnée y sont présenté comme l'écart-type(standard deviation SD) la médiane et les points extrêmes.

On voit d'ailleur que l'écart-type du prix est assez élevé et il est meme plus grand que la moyenne ce qui est rare

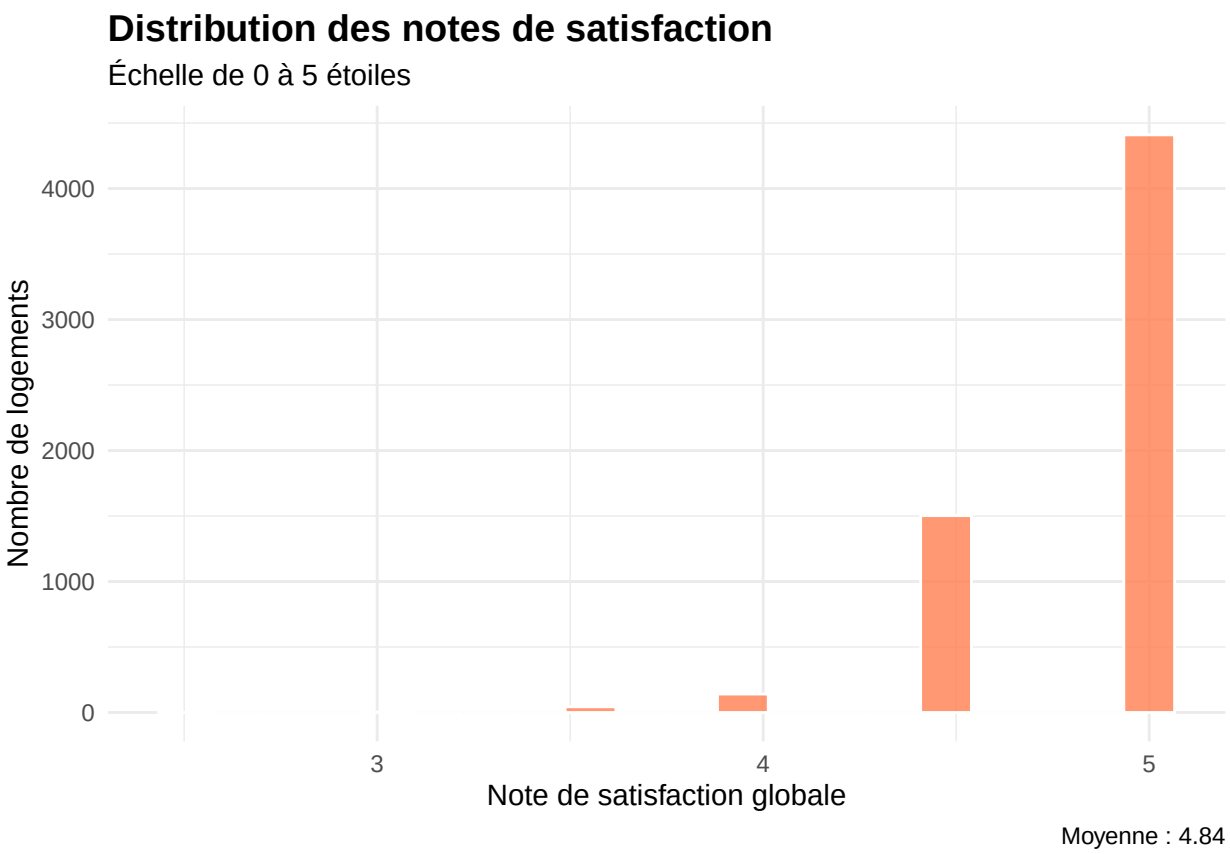
3.2 Distribution des variables clés



Comme dit précédemment il est difficile de voire toutes les valeurs car certaines sont vraiment éloigné et la majorité sont proche de la moyenne comme on peut le voire

avec environ 7000 qui avoisine le prix de 100 euros tout le reste est éparpiller juste derriere et quelque unes sont vers les 4000 euros la nuit

3.3 Distribution de la satisfaction

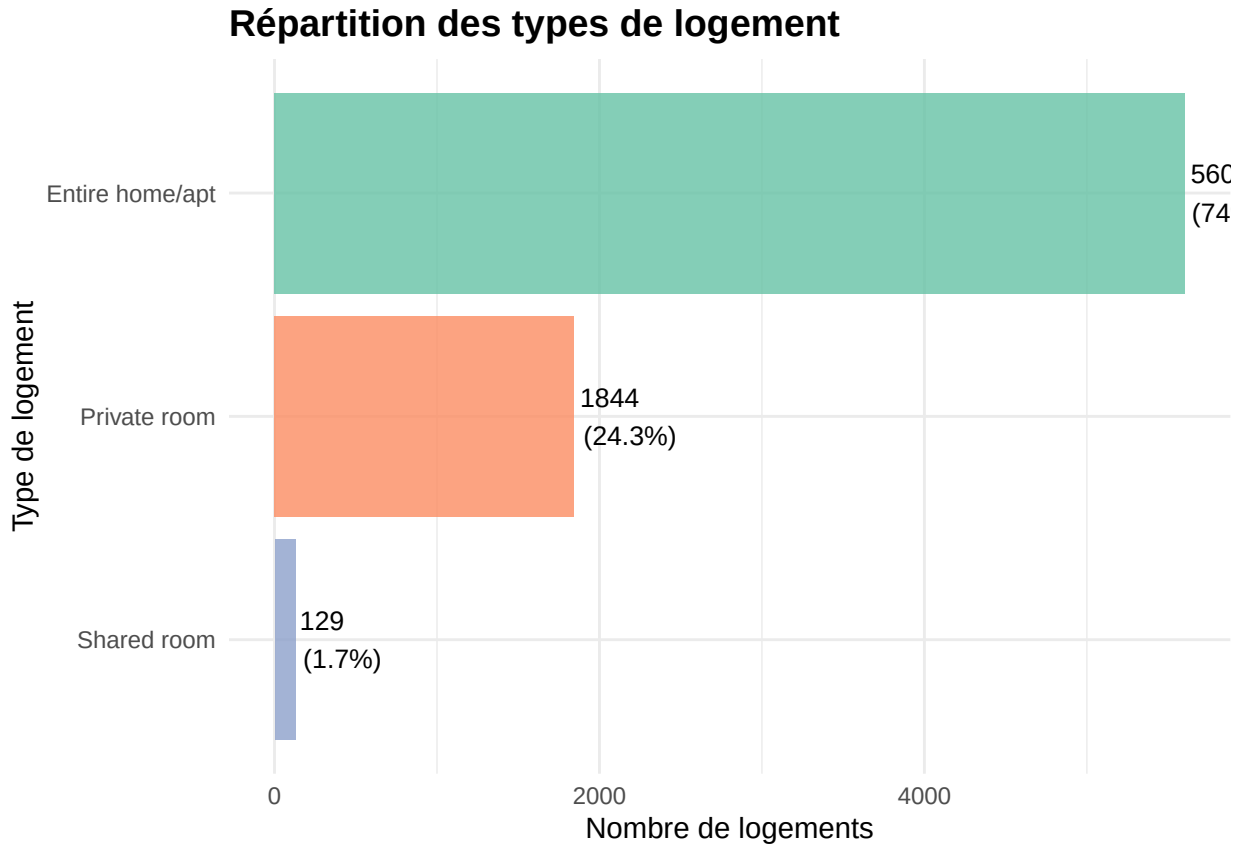


Là encore on a du mal a faire la distinction des minorités sur le diagramme mais ce qui est sur c'est qu'une extrême minorité des logements ont de trés mauvaises notes.

Avec l'écrasente majorité des cas qui se situe entre 4 et 5 étoiles dans les avis leurs moyennes est vraiment bonne en générale.

Cela pose la question : est ce que la note de satisfaction est vraiment un facteur qui afflue sur le prix si la majorité des logements ont une note excellente?

3.4 Distribution des types de logement

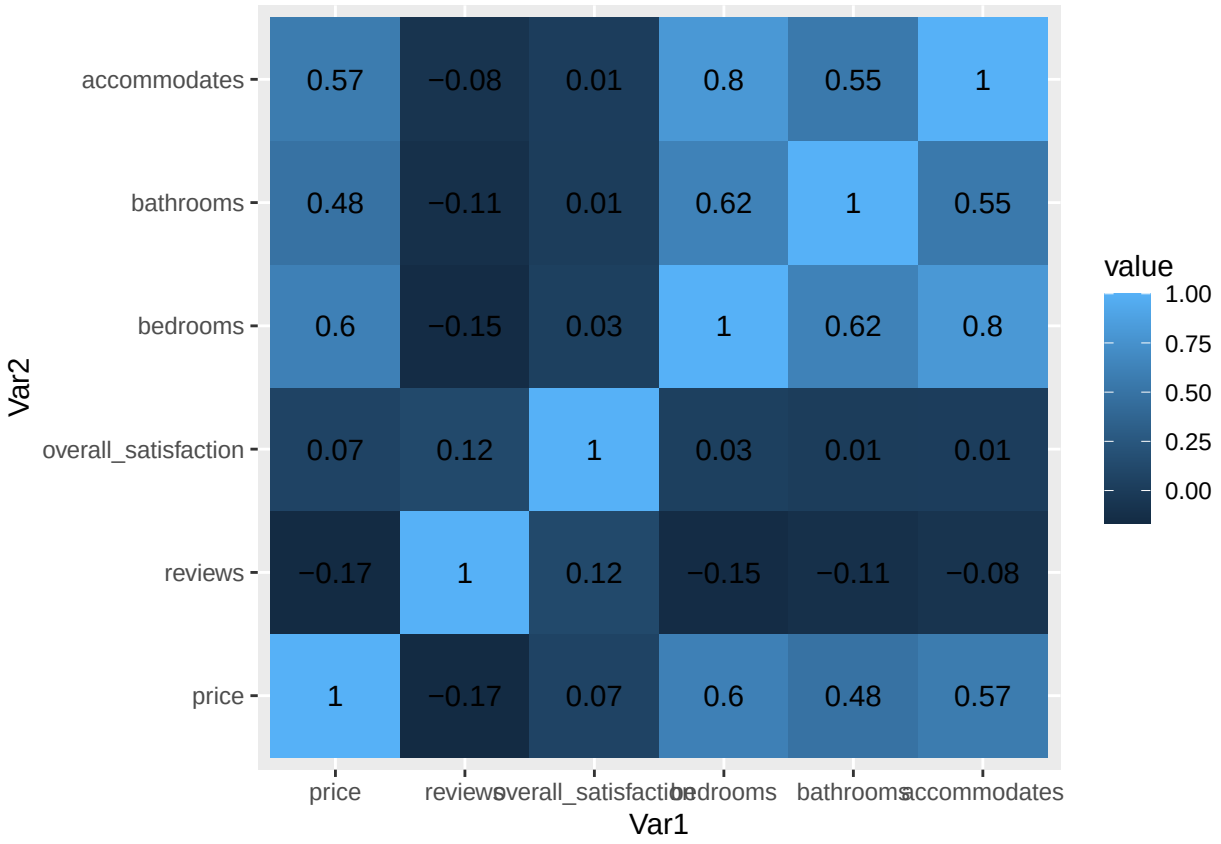


Finalement on peut voire sur ce nouveau diagramme que 74% des airbnb sont des appartements à louer ce qui est en toute logique la base de leurs plateformes de vente.

Avec un nombre minime de chambre partagé et quelques chambres privée encore une fois est ce que le type de chambre afflue sur le prix?

4 Analyse exploratoire bivariée

4.1 Corrélations entre variables numériques



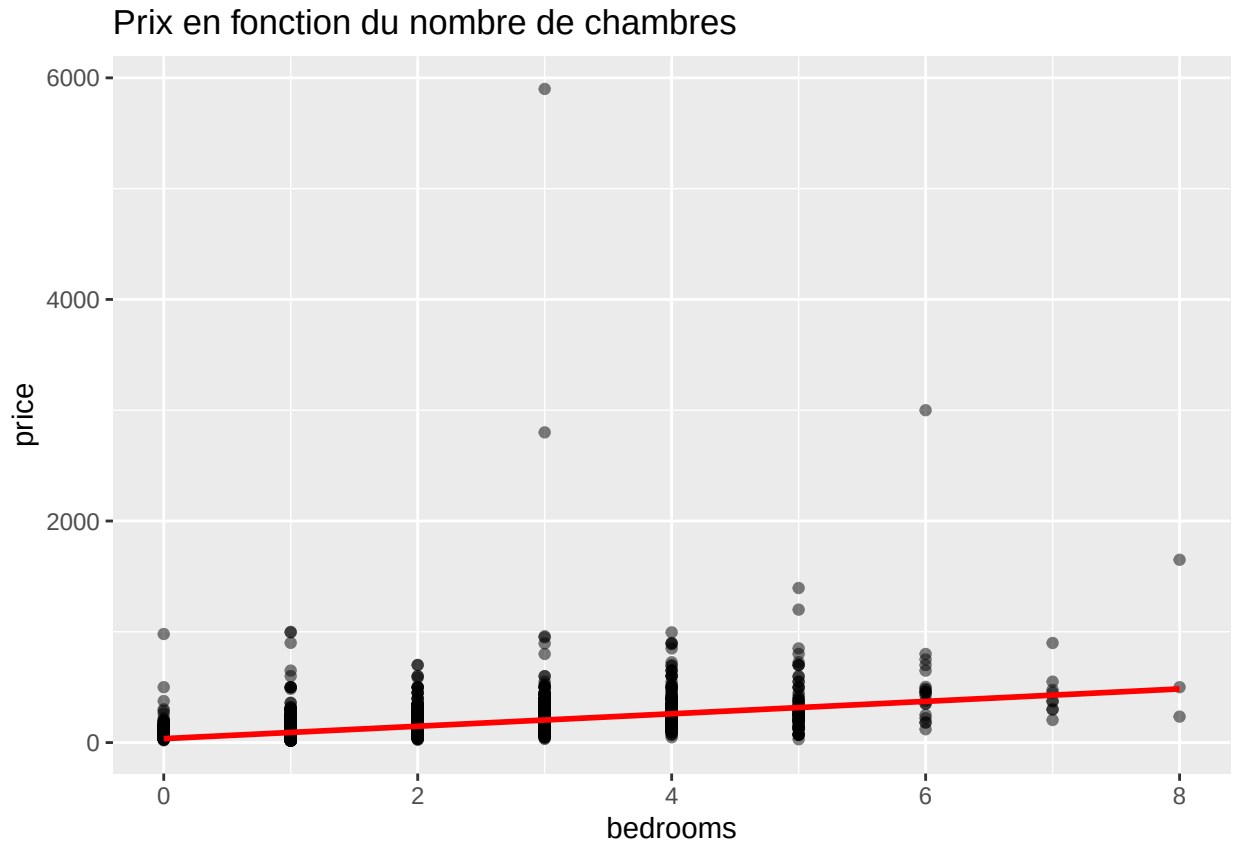
Au départ nous nous posions la question de la corrélations entre les variables. Et bien il extiste une matrice de corrélation qui peut mettre en exergue celle-ci.

Analysons étapes par étapes cette matrice:

- En premier lieu nous voyons que la diagonale des croisements entre meme variables est toujour égale à 1 ce qui correspond à une corrélation parfaite
- Ensuite nous voyons que les nombres assombries sont à faibles coefficients voire des valeurs négatives donc de très faible corrélation
- Finalement on voit certaine cases avec des coéfficients assez haut comme ceux des chambres, salles de bains / toilettes et la capacité d'hébergement

Maintenant que nous avons une idée des principales des variables numériques prenons un exemple et faisons une études pour le nombre de chambre qui parait assez évident

4.2 Visualisation : boite a moustache



Comme on le voit bien ici la tendence montre que le prix augmente avec le nombre de chambre ce qui valide le lien de corrélationentre les deux

On remarque aussi que ce graphique montre d'avantage les valeurs extrêmes comparé aux autres.

5 Modélisation

Qu'est ce qu'une régréssion? Une régression est une méthode statistique qui permet de modéliser la relation entre une variable à expliquer et une ou plusieurs variables explicatives

Ce qui est exactement ce que l'on cherche à faire

5.1 Régression linéaire multiple

Table 3: Résumé du modèle linéaire

term	estimate	std.error	statistic	p.value
(Intercept)	-72.596	14.182	-5.119	0
reviews	-0.111	0.012	-9.118	0
overall_satisfaction	19.337	2.920	6.623	0
bedrooms	26.307	1.491	17.644	0
bathrooms	21.717	1.718	12.640	0
accommodates	7.964	0.596	13.357	0

Analisons maintenant la régréssion qui nous dit:

On voit pour chaque unité de variable ajouté qu'il y a une valeur ajouté en dollar au prix :

- 8 dollar par unité de capacité en plus
- 21.7 pour les salles de bains
- 26 pour les chambres

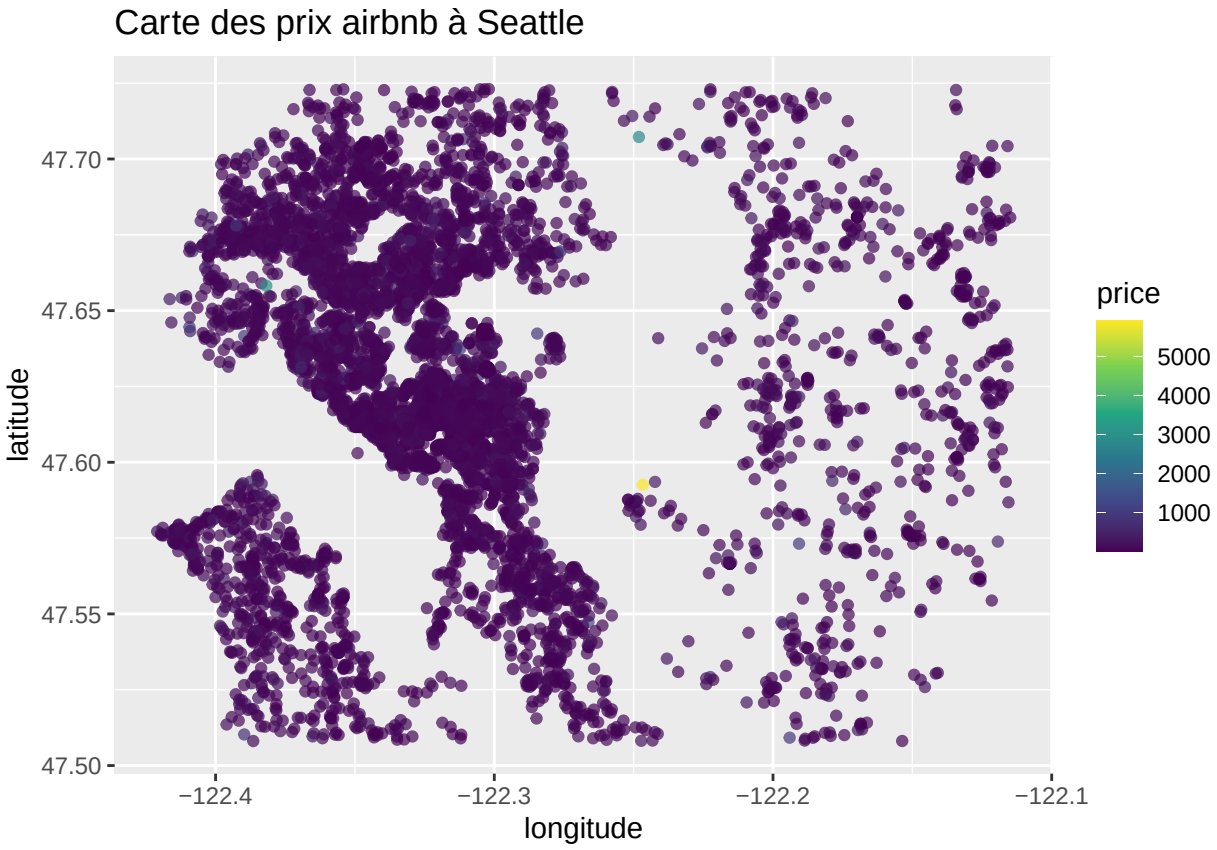
5.2 Régression logarithmique

Table 4: Résumé du modèle logarithmique

term	estimate	std.error	statistic	p.value
(Intercept)	2.915	0.100	29.052	0.000
reviews	-0.001	0.000	-11.805	0.000
overall_satisfaction	0.220	0.021	10.651	0.000
bedrooms	0.140	0.011	13.296	0.000
bathrooms	0.032	0.012	2.632	0.009
accommodates	0.092	0.004	21.906	0.000

6 Analyse géographique

6.1 Cartographie des prix par coordonnées



7 Interprétation et conclusions

7.1 Facteurs influençant le prix

7.2 Limites de l'analyse et recommandations