

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 :	9
5	Modélisation	10
5.1	Régression logarithmique	11
6	Analyse géographique	12
6.1	Cartographie des prix par coordonnées	12
7	Interprétation et conclusions	13

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 observer le lien de corrélation et/ou de causalité entre les paramètres mis à disposition dans le jeu de données. Nous nous attarderons sur les points importants 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 paramètres 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, sûrement celle d'une table propriétaire
- `room_type` : le type de chambre ; on en compte 3 différents (chambre privée, chambre partagée, appartement)
- `adress` : l'adresse de la ville dans laquelle l'Airbnb se trouve
- `reviews` : le nombre d'avis sur l'Airbnb
- `overall_satisfaction` : la note moyenne de l'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 de l'Airbnb
- `latitude` : cet élément représente la latitude de l'Airbnb
- `longitude` : cet élément représente la longitude de l'Airbnb
- `location` : correspond a la version de latitude et de longitude reconnaissable par la base de données 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

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

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

Des moyennes importantes sont aussi visibles, comme celles des reviews qui sont assez élevées, et celle du prix à la nuit qui est de 113 dollars américains.

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

On voit d'ailleurs que l'écart-type du prix est assez élevé, et il est même plus grand que la moyenne, ce qui est rare.

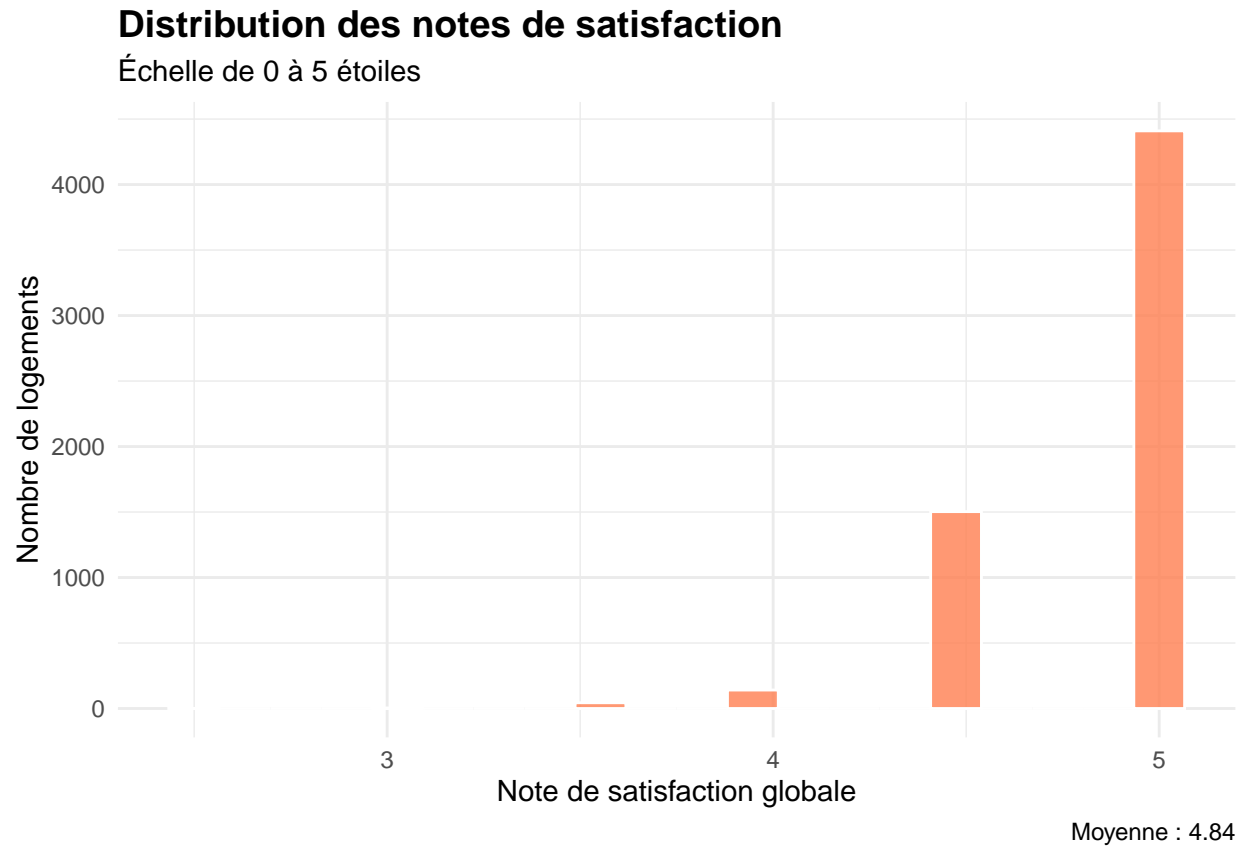
3.2 Distribution des variables clés



Observation : Comme dit précédemment, il est difficile de voir toutes les valeurs, car certaines sont vraiment éloignées et la majorité est proche de la moyenne, comme on peut le voir.

Avec environ 7000 valeurs qui avoisinent le prix de 100 euros, tout le reste est éparpillé juste derrière et quelques-unes sont vers les 4000 euros la nuit.

3.3 Distribution de la satisfaction

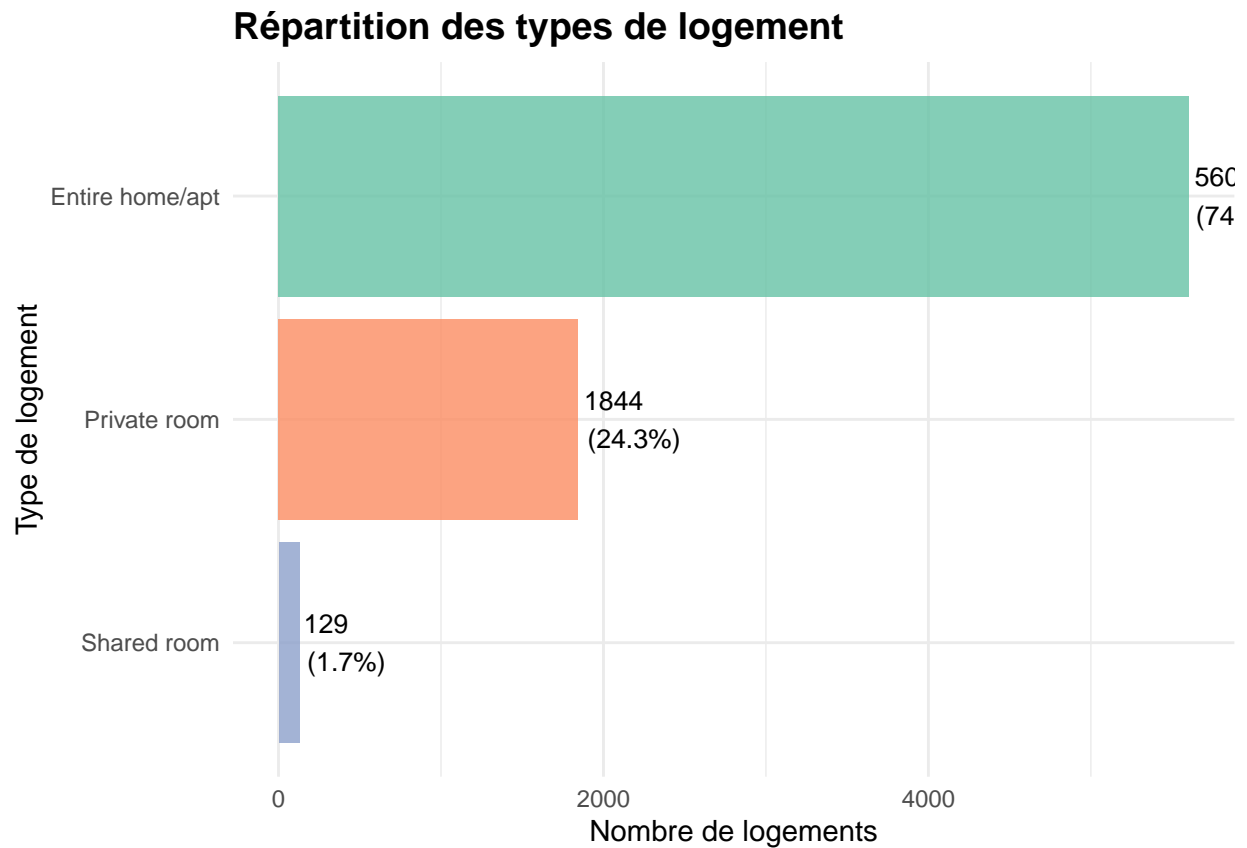


Observation : Là encore, on a du mal à faire la distinction des minorités sur le diagramme, mais ce qui est sûr, c'est qu'une extrême minorité des logements ont de très mauvaises notes.

Avec l'écrasante majorité des cas qui se situe entre 4 et 5 étoiles dans les avis, leurs moyennes sont vraiment bonnes en général.

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

3.4 Distribution des types de logement

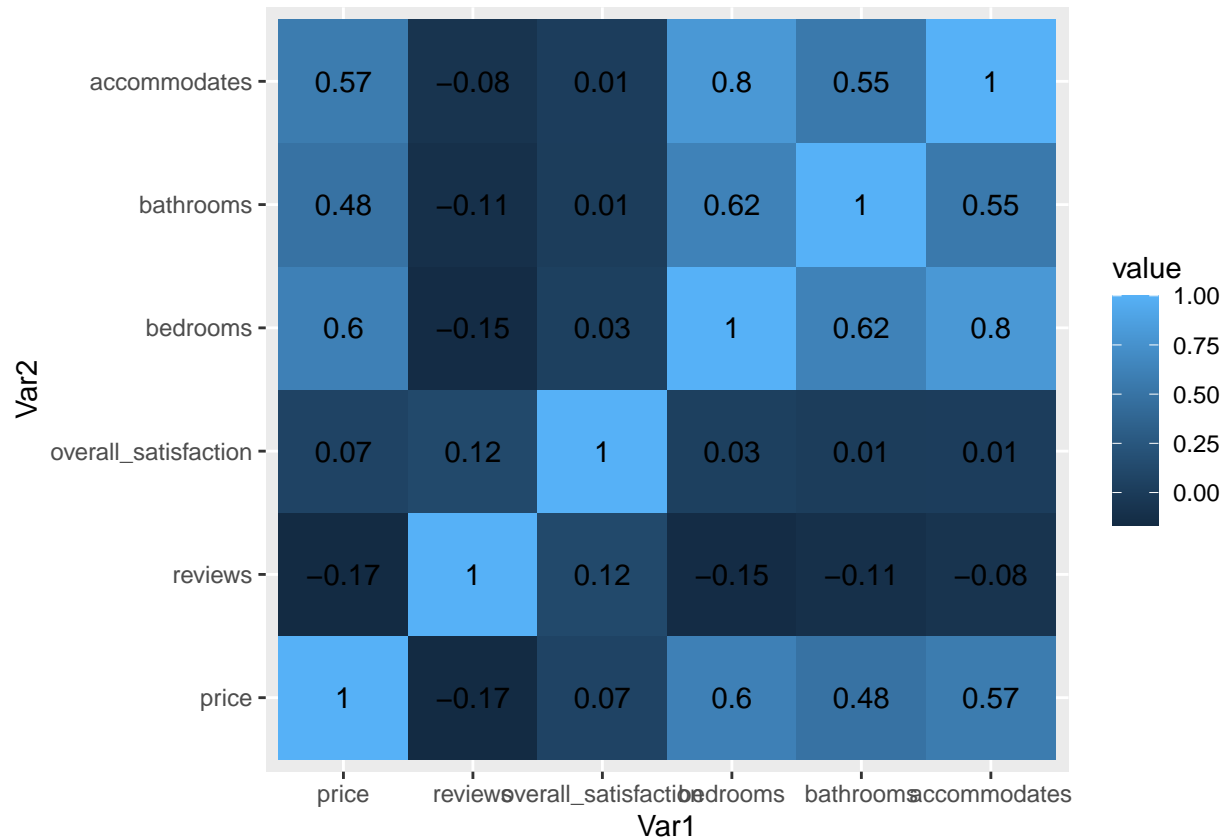


Observation : Finalement, on peut voir sur ce nouveau diagramme que 74 % des Airbnb sont des appartements à louer, ce qui est en toute logique la base de leur plateforme de vente.

Avec un nombre minime de chambres partagées et quelques chambres privées, encore une fois, est-ce que le type de chambre influe sur le prix ?

4 Analyse exploratoire bivariée

4.1 Corrélations entre variables numériques



Observation : Au départ, nous nous posons la question de la corrélation entre les variables. Et bien, il existe une matrice de corrélation qui peut mettre en exergue celle-ci.

Analysons, étape par étape, cette matrice :

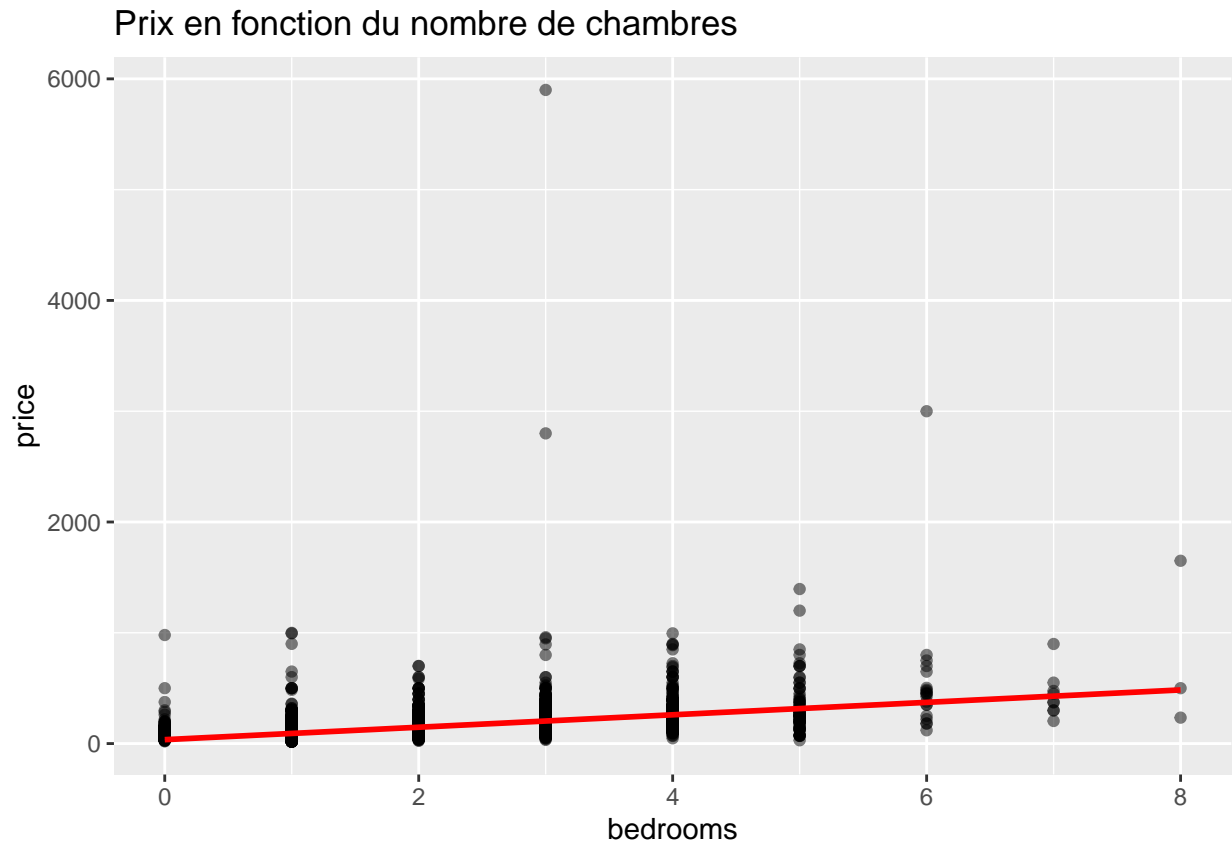
-En premier lieu, nous voyons que la diagonale des croisements entre mêmes variables est toujours égale à 1, ce qui correspond à une corrélation parfaite.

-Ensuite, nous voyons que les nombres assombris sont à faibles coefficients, voire des valeurs négatives, donc de très faible corrélation.

-Finalement, on voit certaines cases avec des coefficients assez hauts, comme ceux des chambres, salles de bains / toilettes et de la capacité d'hébergement.

Maintenant que nous avons une idée des principales variables numériques, prenons un exemple et faisons une étude pour le nombre de chambres, qui paraît assez évident.

4.2 Visualisation :



Observation : Comme on le voit bien ici, la tendance montre que le prix augmente avec le nombre de chambres, ce qui valide le lien de corrélation entre les deux.

On remarque aussi que ce graphique montre davantage les valeurs extrêmes comparé aux autres.

5 Modélisation

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

C'est exactement ce que l'on cherche à faire. ## 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

Observation : Analysons maintenant la régression, qui nous dit :

On voit que pour chaque unité de variable ajoutée, il y a une valeur ajoutée en dollars au prix :
-0,11 \$ pour chaque review en plus

-8 \$ par unité de capacité en plus

-21,7 \$ pour les salles de bains

-26 \$ pour les chambres

-20 \$ pour les unités de satisfaction

5.1 Régression logarithmique

Table 4: Résumé du modèle logarithmique

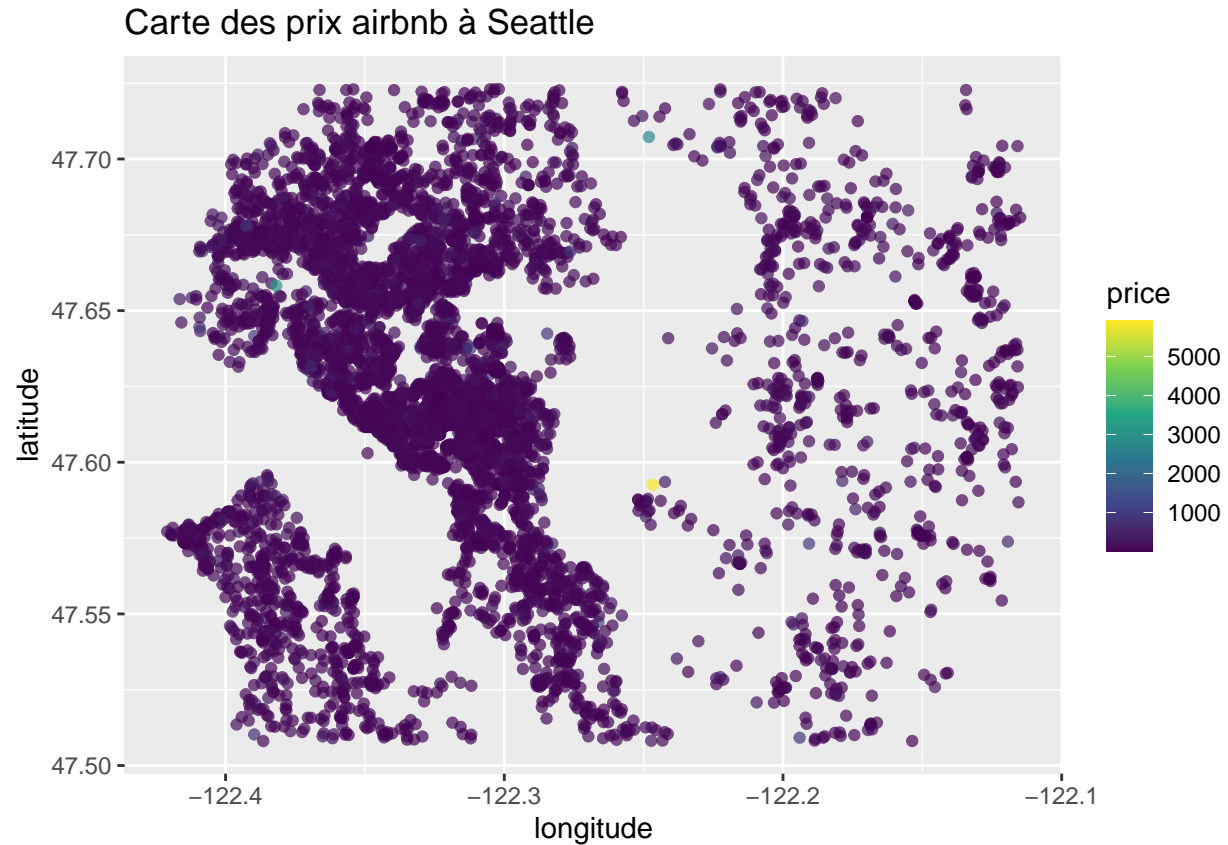
term	estimate	std.error	statistic	p.value
(Intercept)	2.915	0.100	29.052	0.000
reviews	-0.001	0.000	-11.805	0.000
overall_satisfaction	0.220	0.021	10.651	0.000
bedrooms	0.140	0.011	13.296	0.000
bathrooms	0.032	0.012	2.632	0.009
accommodates	0.092	0.004	21.906	0.000

Observation : Pour la régression logarithmique c'est une valeur en % qu'il nous retourne et non une valeur en dollar On voit pour chaque unité de variable ajouté qu'il y a une augmentation du prix en % :

- -1% pour chaque avis en plus
- +22% pour chaque point de satisfaction
- +14% pour les chambres
- +3% pour les salles de bains
- +9% pour la capacité

6 Analyse géographique

6.1 Cartographie des prix par coordonnées



Observation : Maintenant que les variables numériques ont été étudiées, nous pouvons passer aux autres, à savoir la latitude, l'adresse, la longitude et la localisation, qui forment la géographie des Airbnb.

Ce qu'on peut déduire de cette carte, c'est que la concentration est très forte dans la partie ouest de la ville.

On notera qu'il y a une sorte de rivière visible sur l'application qui sépare les deux parties de points. On observe aussi qu'il y a très peu de différences de couleurs, et tout cela à cause des valeurs extrêmes, comme déjà mentionné.

7 Interprétation et conclusions

Conclusion : Notre analyse des données Airbnb de Seattle nous permet de répondre à la problématique initiale concernant les facteurs influençant le prix des locations. Les résultats démontrent que les caractéristiques physiques des logements constituent les déterminants principaux, avec le nombre de chambres comme facteur le plus significatif, suivi par les salles de bain et la capacité d'accueil. La satisfaction des clients, bien que moins déterminante que les caractéristiques structurelles, exerce néanmoins une influence notable sur la fixation des prix. La distribution géographique révèle une forte concentration des offres dans le centre-ville, sans pour autant montrer des disparités de prix marquées entre les différents secteurs. Cette étude confirme ainsi l'importance des attributs tangibles du logement dans la stratégie tarifaire, tout en soulignant la valeur économique de la réputation numérique dans l'économie collaborative.