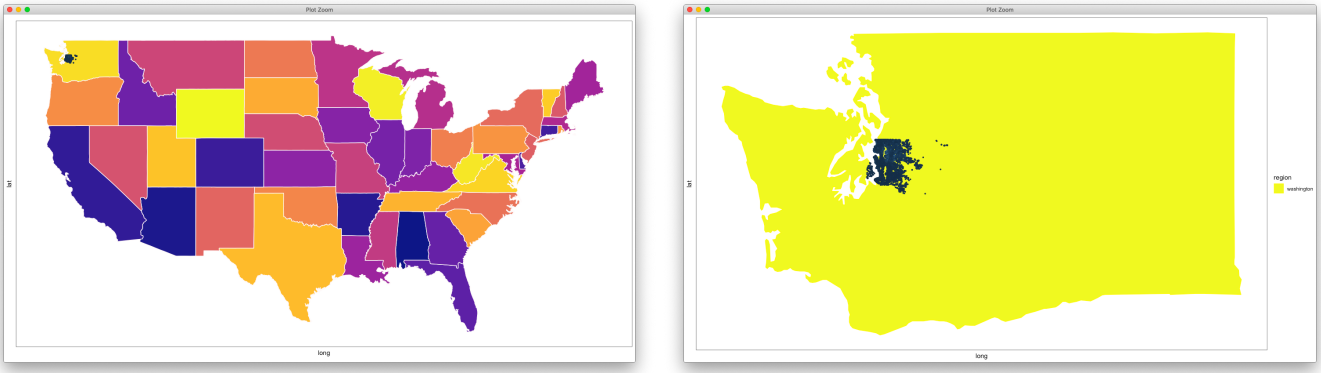
`E2-L3S2` Le prix des logements dans l'État de Washington aux États-Unis

Lola LUBIN et Alexis VIALATTE

2021

Quelles sont les variables qui expliquent le prix des logements dans l'État de Washington aux États-Unis ?

Pour répondre à cette question, nous avons commencé par regarder la base de données `BDD_data.csv` contenant 21613 logements et 21 variables. Nous avons voulu savoir où avaient été recueillies les 21613 données que nous allions analyser. Nous avons enquêté sur la variable `zipcode` et nous avons découvert qu'elle était comprise entre 98001 et 98199 ce qui correspond grossièrement à la ville de Seattle et à ses alentours dans l'État de Washington aux États-Unis. Après avoir regardé la variable `zipcode`, nous avons commencé à regarder les autres variables et nous avons décidé de les conserver, de les transformer ou de les supprimer de nos variables explicatives en sachant que la variable `price` était notre variable expliquée.



Nous avons supprimé la variable `date` de nos variables explicatives car elle n'avait pas d'influence sur le prix des logements. Nous avons conservé les variables `sft_living` et `sft_lot`, correspondant respectivement à la surface intérieure en pieds carrés et à la surface extéieure en pieds carrés, car elles avaient une influence croissante sur le prix des logements. Nous avons également constaté qu'un pied carré intérieur avait plus de valeur qu'un pied carré extérieur ce qui nous a semblé logique. Nous avons conservé les variables `bedrooms`, `bathrooms` et `floors`, correspondant respectivement aux nombre de chambres, de salles de bain et d'étages, car elles avaient une influence croissante sur le prix des logements. Nous avons supprimé les variables `sqft_built`, `sqft_renovated`, `sqft_living15` et `sqft_lot15` car elles étaient plus ou moins équivalentes des variables `yr_built` et `yr_renovated`. Nous avons transformé les variables `yr_built` et `yr_renovated`, correspondant respectivement à l'année de construction et à l'année de rénovation, pour qu'elles correspondent à l'âge depuis sa construction et à l'âge depuis sa rénovation. Nous avons conservé les variables `waterfront`, `view`, `condition` et `grade`, correspondant respectivement à la vue sur la mer, à son point de vue, à son état et à son *design*, pour qu'elles servent à analyser catégoriquement le prix des logements.

Nous avons analysé de manière descriptive les variables `zipcode`, `lat` et `long` car nous les avons supprimé de nos variables explicatives. Nous avons conclu à l'absence de relation suffisamment importante entre les coordonnées géographiques des logements et le prix de ces derniers. Néanmoins, nous avons observé un éclaircissement du nuage de points au centre de Seattle ce qui nous a semblé logique. Nous en avons déduis que plus les logements sont au centre de Seattle et plus le prix de ces derniers est élevé.

Notre première analyse est basée sur des régressions linéaires qui nous ont semblé impératives. Nous avons

analysé le prix en fonction de la surface intérieure et de la surface extérieure en pieds carrés. D'après les différents F, R^2 et les différentes valeurs des constantes et des coefficients, nous avons conservé le modèle log-log sans la surface extérieure en pieds carrés. Nous avons analysé les résidus et toutes les hypothèses nécessaires de ces derniers sont avérées sauf l'hétéroscédasticité. Mais cela ne pose pas de problème car les prix des logements ne sont pas homogènes. Nous pouvons donc conclure que plus la surface intérieure en pieds carrés est élevé et plus le prix est élevé. Par la suite, nous avons essayé de voir si nous pouvions conserver l'homoscédasticité des résidus. Nous avons segmenté notre base de données selon différentes tranches de prix données par l'analyse descriptive de la variable `price` et nous avons conservé l'homoscédasticité si les prix des logements étaient compris entre 321 950 dollars et 450 000 dollars. Malheureusement, le R^2 n'a pas été satisfaisant.

Notre deuxième analyse était basée sur des régressions linéaires en fonction du nombre de chambres, de salles de bain et d'étages. Nous avons observé que la variable `bathrooms` avait l'influence la plus importante. Toutes les hypothèses nécessaires des résultats étaient avérées. Pour les résidus, nous avons eu un problème pour l'hypothèse d'autocorrélation des résidus qui n'était pas avérée. Nous avons encore segmenté notre base de données selon les différentes tranches de prix donné par l'analyse descriptive de la variable `price` et nous avons conservé l'absence d'autocorrélation si les prix des logements étaient compris entre 75 000 dollars et 321 950 dollars. Nous en avons conclu que plus les logement ont de chambres, de salles de bain et d'étages et plus le prix de ces derniers est élevé. Cette analyse est logique puisque plus un logement a de chambres, de salles de bain et d'étages et plus le logement a une surface intérieure en pieds carrés élevée.

Notre troisième analyse était basée sur des régressions linéaires en fonction de la vue sur la mer, du point de vue et du *design* pour les logements les plus luxueux de notre base de données. Nous n'avons conservé que les logements coûtant plus de 645 000 dollars donné par l'analyse descriptive de la variable `price`. Nous avons observé que les variables catégorielles les plus intéressantes étaient celles qui étaient les plus élevées. Toutes les hypothèses nécessaires des résultats et des résidus étaient avérées. Nous en avons conclu que plus les logement sont de grand *stranding* et plus le pris de ces derniers est élevé. Cette analyse est logique puisqu'un logement classe et original aura plus de valeur qu'un logement classique.

Notre quatrième analyse était basée sur des régressions linéaires en fonction de l'état. Nous avons observé que les variables catégorielles les plus intéressantes étaient celles qui étaient les plus élevées. Toutes les hypothèses nécessaires des résultats étaient avérées. Pour les résidus, nous avons eu un problème pour l'hypothèse d'autocorrélation des résidus qui n'était pas avérée. Nous avons une dernière fois segmenté notre base de données selon les différentes tranches de prix donné par l'analyse descriptive de la variable `price` et nous avons conservé l'absence d'autocorrélation si les prix des logements étaient compris entre 75 000 dollars et 321 950 dollars. Nous en avons conclu que plus les logement sont en bon état et plus le prix de ces derniers est élevé. Cette analyse est logique puisqu'un logement insalubre aura moins de valeur qu'un logement neuf ou refait à neuf.

Notre cinquième analyse était basée sur des régressions linéaires en fonction de l'âge de la construction et de l'âge de la rénovation et d'une interaction de ces deux dernières qui nous semblaient liées. Nous avons observé que les variables `life_renovated` et `life_built.life_renovated` étaient égales à 0. Toutes les hypothèses nécessaires des résultats et des résidus étaient avérées. Nous en avons conclu que plus les logements sont récents et plus le prix de ces derniers est élevé.

Notre sixième analyse était basée sur des régressions linéaires en fonction des résultats de nos analyses précédentes dans le but d'avoir une analyse globale. Malheureusement, le R^2 n'étant pas satisfaisant nous a conduis à abandonner cette dernière analyse.

Pour conclure, nos analyses nous ont permis de constater que certains de nos a priori étaient corrects comme la surface intérieure en pieds carrés, le nombre de chambres, de salles de bain, d'étages, la vue sur la mer, le point de vue, l'état, le *desing*, l'âge de la construction et d'autres étaient faux comme la surface extérieure en pieds carrés et l'âge de la rénovation. D'après nos différents résultats, notre conclusion à propos de ce projet est que la variable `sqft_linvg` a l'influence la plus importante dans la relation du prix des logements puis les variables `bedrooms`, `bathrooms` et `floors` et enfin les variables catégorielles `condition` et dans une moindre mesure la variable `life_built`.