Projet Universitaire Régression sur des données réelles

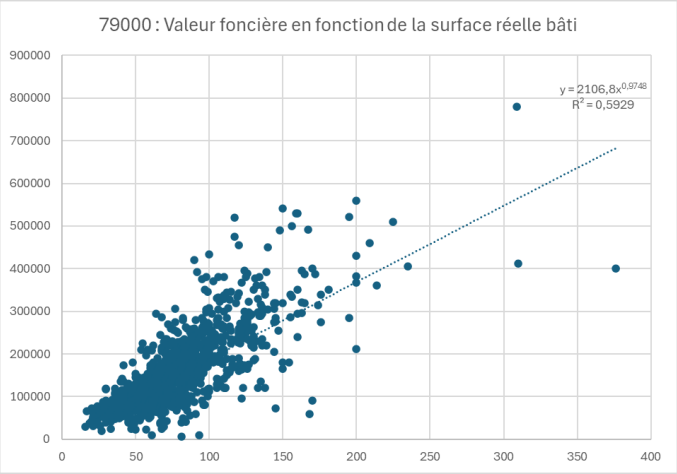
BENOIT Nathaël – GASSE Armand-Valentin

Lors de ce projet universitaire “Régression sur des données réelles”, nous avons pu travailler sur les données de ventes de logements, maisons et appartements, dans le département des Deux-Sèvres en 2023. Un fichier ‘train.csv’ nous a été fourni pour comprendre la structure des données et réaliser des tests dessus, afin de trouver des modèles les plus précis possibles, pour prédire la valeur foncière des logements. Notre but était de créer un modèle pertinent afin de pouvoir prédire la valeur foncière des logements dans le département des Deux-Sèvres en 2024.

Pour développer notre modèle, nous avons d'abord travaillé sur l’ensemble du fichier, en recherchant des relations entre la valeur foncière, la surface réelle bâtie, le nombre de pièces et la surface du terrain. Nous avons rapidement compris que travailler uniquement sur le nombre de pièces était inutile, car cette variable est quantitative discrète, donc non continue. Les variables comme les communes ont été utilisées comme filtre, mais elles ne sont pas prises en compte dans les calculs.

Pour commencer nos recherches, nous nous sommes beaucoup concentrés sur le calcul du SCR pour chaque test effectué, bien qu’il ne soit pas très pertinent. Nous nous sommes ensuite rendu compte qu’il était préférable de nous focaliser sur la création de nuages de points, en utilisant les équations des courbes de régression et le r^2. Cela nous a conduit à effectuer des calculs spécifiques dans le but d’obtenir un r^2 le plus élevé possible. Nous avons testé différents modèles sur l’ensemble des données : un modèle puissance pour la valeur foncière en fonction de la surface réelle bâtie, avec un r^2 de 0.1815, puis un modèle logarithmique, donnant un r^2 de 0.257.

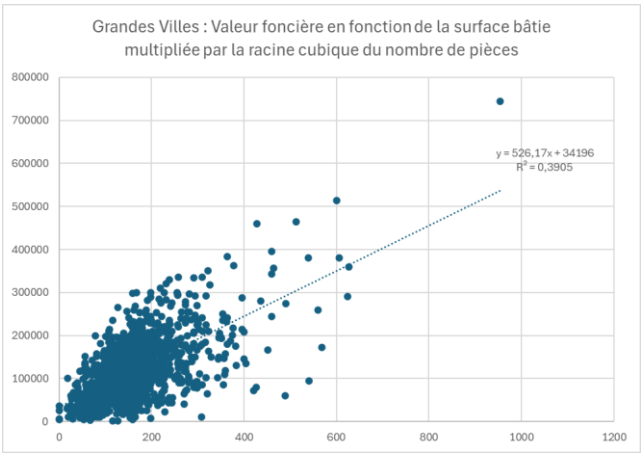
Nous avons vite compris qu’il était difficile de trouver des liens significatifs en travaillant sur l’ensemble du département. Nous avons donc pris la décision de diviser nos données en plusieurs classes, séparant d’abord les appartements des maisons. Pour les maisons, nous avons testé la relation vis-à-vis de la valeur foncière en fonction du produit du nombre de pièces et de la surface du terrain, obtenant un r^2 de 0.1017, puis en fonction du quotient de la surface du terrain et de la surface bâtie, avec un r^2 de 0.0002. Ces tests n’ont alors pas donné de résultats pertinents, ni pour les maisons ni pour les appartements. Ainsi, nous avons décidé de travailler spécifiquement sur Niort, séparément du reste du département. C'est ainsi que nous avons réalisé des tests simples mais efficaces sur la ville de Niort. En effet, en utilisant un modèle puissance pour la valeur foncière en fonction de la surface réelle bâtie, nous avons obtenu un r^2 particulièrement pertinent de 0.5909. Cependant, bien que nous ayons essayé plusieurs



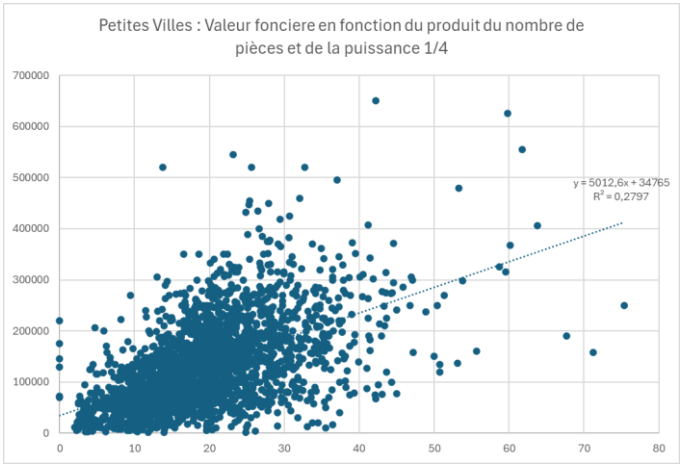
autres approches sur Niort, aucune n'a produit de résultats comparables ou supérieurs à ce r^2. Nous avons alors essayé de rallier l'ensemble des communes ayant pour code postal 79000, afin de tenter d'améliorer notre modèle de prédiction, et on a réussi à obtenir un r^2 de 0.5929 en reprenant le même modèle que celui de Niort. En revanche, pour le reste du département, aucun de nos tests n'a permis d'obtenir un r^2 dépassant 0.3. Le meilleur résultat trouvé était un r^2 de 0.2826, obtenu en étudiant la valeur foncière par rapport à la surface réelle bâtie.

Pour identifier des modèles pertinents dans les villes hors 79000, nous avons entrepris de distinguer celles qui ont vendu plus de 100 logements de celles qui ont réalisé moins de ventes. Les résultats étaient plus intéressants pour les premières, avec des r^2 de 0.3289 et 0.3313, respectivement pour la valeur foncière en fonction de la surface bâtie, et la valeur foncière en fonction du produit de la surface bâtie et du nombre de pièces. Cependant, les villes ayant moins de 100 ventes ont engendré des modèles moins satisfaisants, le meilleur restant basé sur la valeur foncière en fonction du produit de la surface bâtie et du nombre de pièces. Cette approche nous a donc semblé peu idéale, et nous avons donc réfléchi à une séparation par code postal.

Ainsi, nous avons séparé arbitrairement les villes en fonction de leur code postal de la façon suivante : celles avec un code inférieur à 79500 d'un côté, et celles avec un code supérieur de l'autre. Bien que nous ayons obtenu un r^2 acceptable de 0.3381 pour les villes avec des codes supérieurs à 79500, cette méthode manquait de logique et a été rapidement abandonnée. Par la suite, nous avons séparé les grandes villes des petites, en nous basant sur les derniers chiffres du code postal (Les grandes villes ont pour code postal 79100, 79200, ..., et les petites villes ont pour code postal 79110, 79120, ...). Cette méthode a produit des résultats plus qu'intéressants : en utilisant un modèle linéaire pour les grandes villes, nous avons obtenu un r^2 de 0.3905 en étudiant la valeur foncière, par rapport à la surface bâtie multipliée par la racine cubique du nombre de pièces. Ce modèle original a ainsi été retenu pour son efficacité.



Cependant, les tests sur les petites villes ont donné des résultats moins convaincants. Un r^2 d'environ 0.27 a été obtenu, mais ce modèle attribuait systématiquement une valeur de 0 aux appartements. Nous avons donc décidé de séparer les maisons des appartements dans les petites villes. C'est alors qu'en appliquant un modèle linéaire sur les maisons des petites villes (en supprimant toutes leurs valeurs manquantes), nous avons obtenu un r^2 de 0.2797, en comparant leur valeur foncière en fonction du produit du nombre de pièces et de la racine quatrième de la surface du terrain.



Enfin, pour inclure toutes les ventes dans un modèle global, nous avons utilisé un modèle exponentiel pour les appartements des petites villes, modélisant la valeur foncière en fonction du quotient de la surface bâtie par le nombre de pièces. Ce modèle a donné un faible r^2 de 0.2312.

Ne trouvant pas d'autres idées pour améliorer ce r^2, nous avons alors décidé de finaliser notre modèle de prédiction à partir de 4 sous-modèles. Le premier, un modèle puissance sur les villes ayant 79000 pour code postal. Le deuxième, un modèle linéaire sur les grandes villes hors 79000. En troisième, un modèle linéaire sur les maisons des petites villes. Pour finir, un modèle exponentiel sur les appartements des petites villes.

Pour conclure, nous pensons avoir réalisé un travail complet et efficace, grâce à une très bonne répartition des tâches, et avons pu nous familiariser davantage sur la manipulation de données sur Excel. Grâce à ce projet, nous avons aussi pu nous exercer à manipuler R, ce qui nous a permis de nous remémorer les bases, ainsi que de développer et d'approfondir nos compétences sur ce logiciel. Nous avons également le sentiment d'avoir bien organisé notre temps, même si nous avons trouvé que nous disposions d'un temps assez contraignant par rapport à la charge de travail attendu.