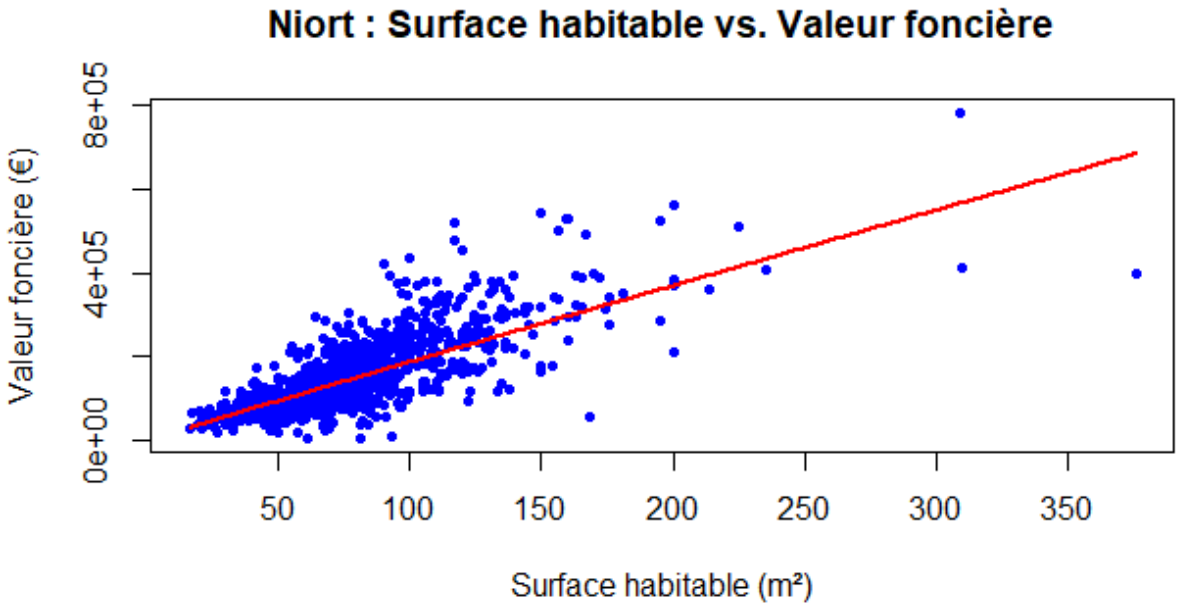
SAE – Régression sur des données réelles

Dans ce projet, l'objectif principal est de prédire le prix de vente de logements (maisons et appartements) situés dans le département des Deux-Sèvres. Pour cela, nous avons utilisé des données immobilières réelles correspondant à des ventes effectuées en 2023 et au premier semestre 2024. Deux fichiers sont mis à notre disposition : un fichier d'entraînement qui contient les caractéristiques des logements (surface, nombre de pièces, localisation, etc.) ainsi que leur prix de vente, et un fichier test qui contient les mêmes informations, mais sans les prix de vente, que nous devons prédire.
Le but du projet est de développer un modèle de prédiction fiable en utilisant les données du fichier d'entraînement, puis d'utiliser ce modèle pour estimer les prix manquants du fichier test. Ce travail sera réalisé entièrement en langage R, sans utiliser de librairies externes. À la fin, nous devrons produire un fichier *prediction.csv* contenant l'identifiant de chaque bien ainsi que son prix estimé. La qualité de nos prédictions sera évaluée par un indicateur basé sur l'écart entre nos estimations et les prix réels.

Au début du projet, nous avons réfléchi à la meilleure manière d'aborder cette tâche. Le fichier d'entraînement comportait des données provenant de plus de 5 000 communes, ce qui compliquait l'intégration de la localisation dans notre modèle de prédiction. Pour simplifier les choses et éviter de nous disperser, nous avons décidé de concentrer notre étude uniquement sur la ville de Niort, qui était bien représentée dans les données. Cela nous a permis de travailler sur un ensemble de données plus homogène et de disposer de suffisamment d'exemples pour construire un modèle solide.
Avant de commencer à implémenter le modèle dans R, nous avons préféré réaliser plusieurs tests sur Excel, un outil plus visuel et plus simple à utiliser pour manipuler rapidement les données. Grâce à des graphiques (nuages de points, courbes de tendance, etc.), nous avons pu mieux comprendre les relations entre les variables. Excel nous a également permis de tester facilement différents types de modèles de régression, en modifiant directement les équations et en observant les résultats.
Pendant cette phase, nous avons exploré différentes combinaisons de variables pour identifier celles qui influençaient réellement les prix. Par exemple, nous avons testé le rapport entre la surface réelle et le nombre de pièces pour estimer la taille moyenne des pièces, ou encore la somme de la surface habitable et de la surface du terrain pour évaluer l'impact global de l'espace sur le prix. Ces essais nous ont permis de mieux cerner les facteurs qui affectaient la valeur foncière des logements.
Après plusieurs essais, nous avons décidé de nous concentrer sur la relation entre la surface réelle des logements et leur valeur foncière, en regroupant maisons et appartements. En effectuant une régression simple, nous avons trouvé un coefficient de corrélation d'environ 0,76, ce qui indiquait une relation relativement forte entre ces deux variables. Nous avons donc choisi de baser notre modèle sur cette corrélation. Toujours sur Excel, nous avons testé trois types de régression : linéaire, logarithmique et puissance. C'est finalement le modèle de régression puissance qui a donné les meilleurs résultats, avec un R^2 plus élevé que les autres. Ce choix a été validé à la fois par les résultats numériques et par ce que nous avons observé visuellement dans les courbes.



Le graphique ci-dessus illustre la relation entre la surface habitable et la valeur foncière pour la ville de Niort. La courbe rouge correspond au modèle de régression puissance que nous avons retenu, qui s'ajuste bien aux points observés.

Une fois que nous avons identifié le modèle qui semblait le plus pertinent, nous avons transféré notre démarche dans R. L'idée était de formaliser et automatiser le processus que nous avions testé sur Excel, afin de pouvoir l'appliquer facilement aux données du fichier test et générer nos prédictions finales.

Pour notre deuxième modèle, nous avons choisi de nous concentrer sur l'ensemble du département des Deux-Sèvres, en excluant la ville de Niort. L'idée était de comparer deux approches différentes : une qui se focalise sur une seule ville (Niort) et une autre qui couvre toutes les autres communes du département.

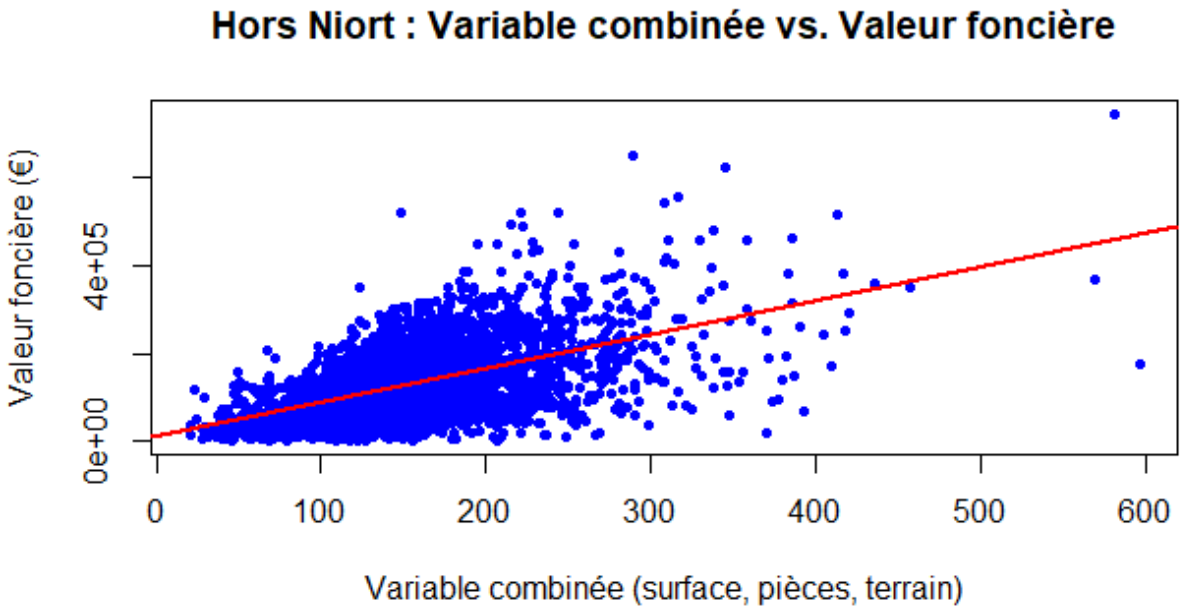
Le but du projet est de trouver les variables qui peuvent expliquer au mieux les prix immobiliers dans le 79, afin de pouvoir prédire ces prix avec notre modèle. Dans cette approche, nous avons voulu aller un peu plus loin dans l'analyse, en intégrant des éléments économiques. Par exemple, nous avons pris en compte l'inflation de 2023. En comparant le taux d'inflation avec le taux nominal de l'année, nous avons obtenu un taux réel d'environ 1,19. Sur cette base, nous avons décidé de multiplier la surface réelle des logements par ce taux réel, dans le but d'ajuster la valeur de l'espace en fonction du contexte économique de l'année.

Ensuite, nous avons examiné l'impact du nombre de pièces sur la valeur du bien. Pour tester cette influence, nous avons choisi de multiplier le nombre de pièces par 9, de manière empirique, pour voir si cela avait un réel effet sur le prix des logements dans notre modèle.

Enfin, pour la surface du terrain, nous avons appliqué une transformation logarithmique. L'idée ici était de réduire l'écart entre les petites et grandes surfaces, et ainsi rendre les valeurs plus faciles à analyser. Une fois cette transformation effectuée, nous avons divisé par 600, car cela correspondait à peu près à la médiane des surfaces de terrain dans notre jeu de données.

Ces ajustements ont donc permis de créer un modèle plus adapté aux autres communes du département, en prenant en compte des aspects économiques et des transformations mathématiques des données brutes.

Le graphique suivant représente la relation entre notre variable transformée (qui combine la surface habitable, le nombre de pièces et la surface du terrain) et la valeur foncière pour l'ensemble du département hors Niort. On peut y observer une tendance cohérente validant la pertinence de notre second modèle.



En conclusion, ce projet nous a permis de mettre en pratique nos connaissances en statistiques et en programmation R sur un cas concret et réaliste. Travailler sur de vraies données immobilières a rendu l'exercice plus motivant et enrichissant. Nous avons appris à nettoyer et structurer les données, à tester plusieurs approches de modélisation, et à interpréter les résultats de manière critique. Même si nous n'avons pas pu explorer toutes les variables comme nous l'aurions souhaité, nous sommes satisfaits du modèle final et du travail réalisé en équipe. Ce projet nous a aussi montré l'importance de bien comprendre les données avant de choisir un modèle, et nous a donné envie d'approfondir nos compétences dans l'analyse de données et la modélisation prédictive.

ERRADDAOUI Yasmine

GARDERE Théo