**Estimation de la valeur foncière des logements dans les Deux-Sèvres (79)**

**Introduction**

Dans le cadre d’un projet de première année de BUT Sciences des Données, nous avons réalisé une étude visant à estimer la valeur foncière de logements situés dans le département des Deux-Sèvres (79). L’objectif était de proposer une modélisation simple mais pertinente à partir de données brutes, en combinant l’exploitation d’un fichier Excel avec l’utilisation de RStudio pour appliquer des modèles de prédiction.

Le jeu de données fourni contenait plusieurs variables explicatives essentielles à la compréhension du marché immobilier local, telles que :

* La ville dans laquelle se situe le bien,
* Le type de logement (maison ou appartement),
* La surface réelle du logement (en m²),
* La surface du terrain (en m²),
* Le nombre de pièces,
* Et enfin, la valeur foncière (en euros), qui constitue la variable cible de notre étude.

**Démarche de recherche**

**1. Création d’une variable prédictive pertinente**

Dans un premier temps, nous avons construit une variable synthétique permettant de prédire approximativement la valeur d’un bien immobilier à partir des données disponibles. Cette variable est construite en combinant des informations sur la surface du terrain, le nombre de pièces et la ville de localisation du bien.

L’idée a été la suivante :

* **Création de la variable explicative Surface.terrain \* Nombre.pieces.principales** :
  + Une nouvelle variable a été créée en multipliant la **surface du terrain** par le **nombre de pièces principales** pour chaque bien immobilier. Cela permet de capturer l'effet combiné de ces deux facteurs sur la valeur foncière du bien.
  + Cette nouvelle variable a ensuite été utilisée pour construire des modèles de régression linéaire et effectuer des prédictions.

**2. Application des coefficients et prédiction de la valeur foncière**

Une fois les coefficients obtenus pour les logements à Niort, ces derniers ont été appliqués aux données de notre échantillon de test (comprenant des logements à Niort et hors Niort) afin de prédire la valeur foncière des logements en fonction de leur surface et du nombre de pièces.

Pour chaque bien immobilier, nous avons utilisé la formule suivante pour calculer la valeur foncière prédite :

Valeur foncière prédite =a×Surface.terrain+b×Nombre.pieces.principales

L’objectif de ce modèle était de fournir une estimation de la valeur foncière en fonction de caractéristiques mesurables et disponibles dans les données, tout en prenant en compte les différences entre les maisons et appartements situés à Niort et ceux situés ailleurs.

* **Pour les maisons situées à Niort et en dehors de Niort**, nous avons élargi la logique :
  + En plus de la **surface réelle** et du **nombre de pièces**, nous avons également tenu compte de la **surface du terrain**, car en zone rurale ou périurbaine, cette surface influe notablement sur le prix final
  + Ce choix de construction d’une variable prédictive repose sur une logique économique : la valeur d’un bien immobilier est généralement fonction de sa taille, de son usage (maison ou appartement) et de son emplacement (urbain ou non). Intégrer le terrain pour les maisons hors Niort ajoute une dimension réaliste au modèle.

**3. Découpage du jeu de données par segments**

Afin d’améliorer la précision des prédictions, nous avons segmenté le jeu de données en plusieurs sous-populations homogènes. Cette méthode permet d’ajuster plus finement le modèle aux spécificités de chaque type de bien.

Nous avons ainsi créé huit segments en fonction de :

* Le type de logement (maison ou appartement),
* La localisation (à Niort ou hors Niort),
* Les quartiles de la **surface du terrain** (uniquement pour les maisons).

Les segments sont donc les suivants :

1. Appartements à Niort
2. Appartements hors Niort
3. Maisons à Niort, terrain < 283 m²
4. Maisons à Niort, terrain ∈ [283 ; 578[
5. Maisons à Niort, terrain ∈ [578 ; 810[
6. Maisons à Niort, terrain > 810 m²
7. Maisons hors Niort, terrain < 283 m²
8. Maisons hors Niort, terrain ∈ [283 ; 578[
9. Maisons hors Niort, terrain ∈ [578 ; 810[
10. Maisons à Niort, terrain > 810 m²

Ce découpage par quartiles nous permet de contrôler la variabilité dans les groupes et de capter des tendances spécifiques à chaque configuration de bien.

**4. Modélisation linéaire dans Excel**

Dans chaque feuille de calcul, nous avons modélisé la valeur foncière en utilisant une régression linéaire simple basée sur notre variable prédictive. Nous avons appliqué la formule **DROITEREG** dans Excel pour obtenir les coefficients **a** et **b**.

**5. Évaluation de la précision des prédictions**

Pour mesurer l’efficacité de notre modélisation, nous avons calculé, pour chaque feuille, l’écart moyen absolu entre les valeurs foncières réelles et les valeurs prédites :

=somme(abs(valeurs predites – valeurs foncieres)

Ce choix de métrique repose sur sa lisibilité et sa robustesse : il permet de quantifier l’écart moyen sans être influencé par d’éventuelles valeurs extrêmes (contrairement à la variance). Cette évaluation nous a aidés à ajuster les coefficients au m² utilisés dans les formules, dans le but d’optimiser l’écart moyen et de rapprocher au mieux les prédictions des valeurs réelles.

**Modèle retenu**

Le modèle final retenu repose sur une formule de type linéaire propre à chaque segment de bien, avec des coefficients calculés de manière empirique à partir de la combinaison **surface\*nombre de pièces**. Ce choix est pertinent pour plusieurs raisons :

* **Simplicité** : Le modèle est simple à interpréter et à mettre en œuvre sur Excel, ce qui le rend facilement accessible et reproductible.
* **Adaptabilité** : Grâce au découpage segmenté, le modèle s’adapte à chaque type de bien, en tenant compte des variations locales et des caractéristiques spécifiques des biens immobiliers.

En outre, une fois cette modélisation réalisée sur Excel, nous avons transposé l’ensemble du travail sur **RStudio**. Cette étape a permis de valider la cohérence de notre méthode tout en exploitant un environnement statistique plus rigoureux, offrant plus de souplesse et de possibilités pour affiner nos résultats.

L’utilisation de RStudio nous a permis d'isoler les différents groupes et de vérifier la pertinence des coefficients obtenus avec Excel. Nous avons ainsi comparé les prédictions faites dans Excel et celles produites par R, et les résultats ont confirmé la fiabilité de notre approche.

**Conclusion**

Ce projet nous a permis de combiner l’analyse de données, la logique économique et la modélisation statistique dans un cadre réaliste. Grâce à des outils simples (Excel et RStudio), nous avons appris à :

* **Créer des variables pertinentes** à partir de données brutes, en combinant des informations facilement accessibles (surface, nombre de pièces, emplacement).
* **Segmenter un jeu de données de manière intelligente**, en créant des sous-populations homogènes pour ajuster les modèles aux spécificités de chaque type de bien immobilier.
* **Modéliser les relations linéaires** et interpréter les coefficients, afin de prédire la valeur foncière en fonction des caractéristiques mesurables des logements.
* **Évaluer la qualité de nos prédictions**, en utilisant des métriques simples comme l'écart moyen absolu pour ajuster les coefficients et optimiser la précision du modèle.

La transition d’Excel vers R nous a permis de mieux comprendre les forces et limites de chaque outil, et d’anticiper les perspectives d’amélioration, telles que