Prédiction de la valeur foncière des biens immobiliers

# 1. Introduction

Le projet avait pour objectif de prédire la valeur foncière des biens immobiliers dans le département des Deux-Sèvres pour les années 2023 et 2024. Les données étaient fournies sous forme de deux fichiers CSV : le fichier 'train', contenant des informations sur les ventes de biens, et le fichier 'test', qui contient les mêmes variables mais sans la variable cible, à savoir le prix de vente. L’objectif était de construire un modèle à partir des données d’entraînement et d'utiliser ce modèle pour prédire les prix des biens dans le fichier de test.  
Pour atteindre cet objectif, une démarche en plusieurs étapes a été suivie, incluant l'exploration des données, la création de variables pertinentes, ainsi que le choix et l'évaluation du modèle le plus adapté pour prédire la valeur foncière.

# 2. Méthodologie

## Exploration des données

Les données fournies comportaient plusieurs variables, telles que la surface habitable, le type de bien (maison ou appartement), et des informations géographiques telles que la localisation du bien. Une première étape a consisté à analyser ces variables pour identifier les plus pertinentes pour la prédiction du prix de vente. Pour ce faire, nous avons notamment étudié le coefficient de corrélation (un indicateur qui permet de voir si une variable a une influence sur une autre)entre chaque variable explicative et la variable cible (le prix de vente), ce qui nous a permis de repérer les variables ayant le plus d'influence sur le prix, et donc de les prioriser dans la modélisation.

## Traitement des valeurs manquantes et extrêmes

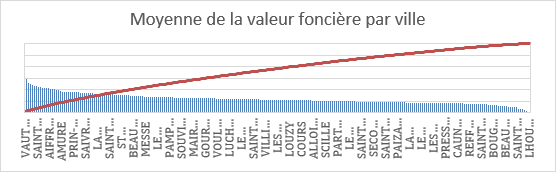
Afin de traiter les valeurs manquantes, nous avons opté pour un remplissage par zéro. Bien que cette approche ne soit pas idéale dans tous les cas, elle nous a semblé suffisamment robuste pour ce projet. Ce choix a été retenu après avoir testé d’autres méthodes comme le remplissage par la moyenne ou la médiane, mais celles-ci faussaient davantage les résultats dans notre cas. Quant aux valeurs extrêmes, nous avons initialement envisagé de supprimer les 2 % des valeurs les plus basses et les plus hautes. Cependant, après réflexion, nous avons choisi de les conserver, estimant que de telles valeurs pourraient également être présentes dans le fichier de test, ce qui permettait de garder une certaine cohérence.

## Création de variables dérivées

Afin de rendre le modèle plus précis, nous avons créé des variables dérivées, comme le prix moyen au m² par ville, en faisant la distinction entre les maisons et les appartements. Pour les appartements uniquement, nous avons ajouté une dimension supplémentaire en prenant en compte le nombre de pièces. Cela a permis d'affiner davantage les estimations du prix au m² par ville. En revanche, pour les **maisons**, cette information s’est révélée moins utile, voire perturbante : lorsque nous l'avons intégrée dans le modèle, la qualité des prédictions diminuai

## Analyse des graphiques

À ce stade, quelques graphiques ont été générés pour mieux comprendre la répartition des données et appuyer notre choix de variables explicatives :



1. Répartition des prix de vente par ville : Ce graphique montre la variation des prix selon la localisation des biens dans les Deux-Sèvres, illustrant l’importance de la localisation dans la détermination du prix. Cela justifie le choix de calculer un prix moyen au m² spécifique à chaque ville.

2. Évolution des prix entre 2023 et 2024 : Un autre graphique a été créé pour comparer l’évolution des prix de vente entre 2023 et 2024. Cela nous a permis d’identifier un taux d’évolution spécifique pour ajuster nos prédictions en fonction de l'année de vente. Cette différence de prix justifie l'inclusion d’un coefficient d'ajustement dans le modèle final.

# 3. Modèle choisi

Après avoir exploré différentes approches, nous avons opté pour un modèle de régression linéaire. Bien qu’il soit relativement simple, ce modèle permet de capturer les relations linéaires entre les variables explicatives (comme le prix au m², la localisation et l’année de vente) et la variable cible, à savoir le prix de vente des biens.  
Le choix de ce modèle a été motivé par sa simplicité et sa transparence, ce qui nous a permis de mieux comprendre l’impact de chaque variable sur le prix foncier. De plus, la régression linéaire s’est avérée être suffisamment robuste pour ce type de prédiction, surtout compte tenu des limitations liées à l’absence de nombreuses variables explicatives.

Les variables retenues dans le modèle final étaient les suivantes :

• Prix au m² par ville.  
• Surface habitable et nombre de pièces pour les appartements.  
• Année de vente (2023 ou 2024).  
• Localisation (avec la distinction entre maisons et appartements).

# 4. Limites du modèle

## Manque de variables explicatives

Le principal obstacle à la précision du modèle a été le manque de variables explicatives. Par exemple, nous ne disposions pas de données sur l’âge du bien, l’état (rénové, à rénover), ou encore la présence de caractéristiques supplémentaires comme un garage ou une piscine. Ces éléments jouent un rôle majeur dans la détermination du prix foncier, et leur absence a limité la capacité du modèle à être totalement précis.

## Traitement des données manquantes et extrêmes

Le choix de remplir les valeurs manquantes par des zéros a permis de maintenir un certain niveau de simplicité dans le traitement des données. Cependant, cette approche peut entraîner une distorsion des prédictions dans certains cas. En outre, bien que nous ayons conservé les valeurs extrêmes, leur présence dans les données a également pu affecter la performance du modèle.

# 5. Conclusion

Ce projet a permis de développer un modèle de régression pour prédire les prix fonciers dans les Deux-Sèvres en 2023 et 2024. Grâce à l’analyse des variables et à la construction de variables dérivées, nous avons pu obtenir un modèle raisonnablement robuste malgré les limitations du jeu de données. Le modèle final repose sur des hypothèses simples mais efficaces, en particulier en ce qui concerne la localisation, le type de bien et l’évolution des prix entre les deux années. Cependant, le manque de variables explicatives cruciales comme l’âge et l’état des biens nous empêche de prédire avec une précision parfaite.  
En dépit de ces limitations, ce travail a fourni une bonne compréhension des techniques de nettoyage des données, de modélisation prédictive et des choix méthodologiques nécessaires dans un contexte d’analyse de données réelles.