SAÉ - Régression sur des données réelles

- Vous avez à votre disposition un jeu de données concernant les ventes immobilières d'appartements et de maisons en 2023 et au premier semestre 2024 dans les Deux-Sèvres. Le jeu est scindé de 2 fichiers csv :
    - train : contient le prix de vente et des variables donnant des informations concernant chaque logement ;
    - test : contient les mêmes variables que train mais pas le prix de vente.

Votre travail consiste à utiliser le jeu d'entrainement (train) pour déterminer le meilleur modèle de prédiction du prix de vente. Vous appliquerez alors votre modèle pour prédire le prix de vente des logements présents dans le fichier test.

Vous devez me rendre :

- Un texte bien structuré (environ deux pages hors illustrations) avec :
    - Une introduction présentant notamment le projet ;
    - Votre démarche de recherche pour élaborer votre modèle. Cette étape passe par l'étude de la variable cible et le croisement de cette variable avec les différentes variables explicatives, les variables qualitatives notamment. Vous pouvez enrichir votre texte avec des graphiques qui vous ont été utiles pour orienter le choix de votre modèle ;
    - L'exposition claire du modèle retenu ;
    - Une conclusion dans laquelle vous pourrez notamment donner votre sentiment sur le travail fourni.
- Un fichier nommé "prediction.csv" au format CSV2 contenant 2 variables : la variable id du jeu test et une variable Valeur.fonciere contenant vos prédictions ;
- Le code R ayant permis de générer ce fichier à partir des 2 fichiers fournis (train et test). Ce code doit commencer par l'importation des fichiers train et test et se terminer par l'export du fichier prediction. **Attention : ni accent, ni espace dans les noms des fichiers, des objets (data.frame) et des variables.**

Le code R ne doit pas utiliser de librairies externes. Il ne doit pas utiliser les fonctions 'lm' et 'predict' !

L'évaluation portera sur :

- La méthodologie utilisée pour élaborer le modèle et la profondeur des recherches menées

- La complexité du modèle retenu
- La qualité et la clarté de votre texte explicatif
- La clarté et l'efficacité de votre code R permettant de créer le modèle que vous avez finalement retenu
- La précision de vos prédictions qui se fera au moyen de la métrique «somme des carrés des résidus". Un classement des meilleures prédictions sera effectué.

Environ la moitié des points est attribuée au rapport.

Vous travaillerez par groupes de deux - Bon courage !