

**Rapport de projet : SAÉ - Régression sur des données réelles**

Pour ce projet d’analyse des valeurs de ventes immobilières à Paris au premier semestre 2023, nous avons eu à notre disposition un ensemble de données comprenant des informations sur les valeurs foncières d’appartements et de maisons. Les données sont réparties dans deux fichiers CSV distincts : « train » et « test ».

Le fichier « train » constitue le jeu de données principal sur lequel nous effectuons nos démarches de recherches. Il contient des détails sur chaque logement vendu, y compris le prix de vente (la valeur foncière), ainsi que diverses autres variables telles que la surface ou le nombre de pièces du logement. Le fichier « test », ne comporte pas la colonne des valeurs foncières.

L’objectif du projet était donc d’utiliser les informations fournies dans le fichier « train » pour construire un modèle prédictif des prix de vente des logements répertoriés dans le fichier « test ».

\*

\* \*

La démarche de recherche pour élaborer le modèle de prédiction du prix de vente des biens immobiliers s’est déroulée sur plusieurs séances de travail. Lors des premières séances, nous avons importé le fichier CSV « train » dans Excel pour une première exploration et manipulation des données. Cette étape nous a permis de mieux comprendre la structure des données et de repérer d’éventuelles valeurs aberrantes ou manquantes.

Par la suite, nous avons entrepris d’explorer différentes combinaisons de variables et d’appliquer différents types de modèles pour prédire le prix de vente des logements. Nous avons d’abord souhaité tester le lien entre la variable surface et d’autres variables, car il s’agit de la variable qui nous a semblé la plus significative. Nous avons l’application de modèles linéaires, exponentiels, logarithmiques et de puissance, en ajustant les paramètres pour obtenir les meilleurs résultats possibles. C’est-à-dire que nous avons sélectionné les variables nous semblant les plus pertinentes pour prédire la valeur foncière des biens immobiliers. Par exemple, nous avons commencé par essayer de trouver un lien entre la surface du logement et sa commune de localisation, ou encore un lien entre la surface et le nombre de pièces.

À chaque étape de combinaison de deux variables, nous avons tenté d’évaluer la performance des modèles en utilisant l’outil qu’est le coefficient de corrélation afin de savoir si le lien entre chaque couple de variables était suffisamment élevé pour être considéré par la suite. Cela nous a donc permis d’identifier les modèles les plus prometteurs à savoir le modèle linéaire et le modèle logarithmique.

De plus, nous nous sommes interrogés sur les variations de prix de vente selon les arrondissements de Paris, car il est évident que la valeur foncière des biens immobiliers n’est pas équivalente dans les arrondissements les plus riches et ceux considérés comme les moins chers.

Avant de valider définitivement le choix de notre modèle, nous avons effectué un tri du jeu de données de base (« train.csv ») afin d’éliminer les données aberrantes qui pourraient fausser les analyses, et de vérifier si les modèles correspondaient bien au jeu de données. Pour ce faire, nous avons déterminé plusieurs contraintes. Tout d’abord, nous avons effectué un filtrage basé sur les valeurs extrêmes des prix de vente des logements en supprimant les données correspondant aux 10 % des valeurs foncières les plus basses, ainsi que les données correspondant aux 10 % les plus élevés. Ensuite, nous avons éliminé les observations présentant des valeurs aberrantes en termes de surface de logement, c’est-à-dire ceux dont la surface était inférieure à 9 mètres carrés, cette surface étant le minimum légal. De plus, nous avons filtré les données pour exclure les logements dont la valeur au mètre carré dépassait 40 000 euros, ce qui nous a permis de nous concentrer sur des biens immobiliers plus cohérents avec le marché parisien.

En appliquant ces filtres successifs, nous avons pu nettoyer le jeu de données de base et nous assurer de la qualité des données exploitées dans notre processus. Cela nous a donc permis de travailler avec un ensemble de données plus fiable et représentatif, propice à l’élaboration d’un modèle précis.

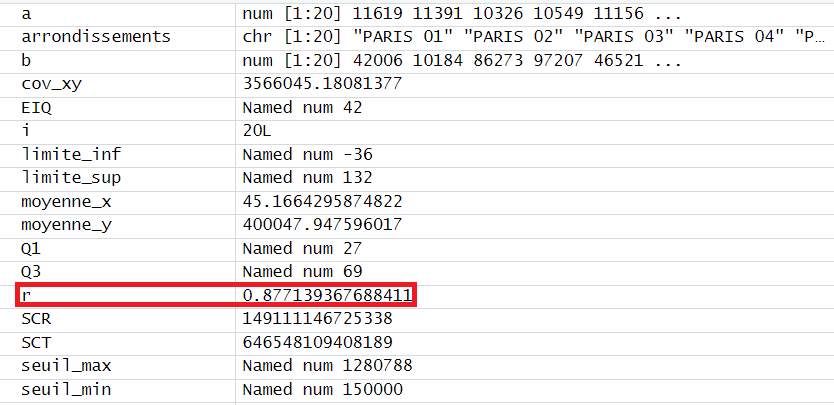
Le modèle pour lequel nous avons opté est un modèle linéaire, car il nous semblait le plus cohérent. Ce modèle repose sur une relation linéaire entre la valeur foncière des biens immobiliers et leur surface réelle. Pour construire ce modèle, nous avons enchaîné plusieurs étapes.

Tout d’abord le nettoyage des données. Nous avons commencé par trier par ordre croissant des valeurs foncières les données du fichier « train.csv »

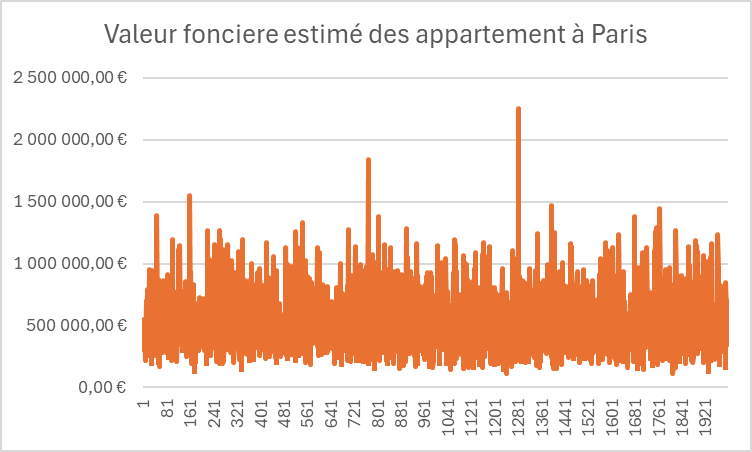
Puis, pour filtrer les données, nous avons appliqué quelques conditions pour éliminer les données aberrantes du fichier « train.csv ». Ce sont celles mentionnées dans le paragraphe sur les démarches de recherche.

Pour construire le modèle final, nous avons décidé de se pencher sur un modèle linéaire pour chaque arrondissement de Paris. Pour ce faire, nous avons calculé les coefficients de régression, a et b, à partir des données filtrées de chaque arrondissement. Ces coefficients ont été déterminés en utilisant la méthode des moindres carrés, où « a » représente la pente de la droite de régression et « b » l’ordonnée à l’origine.

Nous avons ensuite évalué la performance de notre modèle en calculant le coefficient de corrélation linéaire, r, entre les valeurs foncières prédites et celles observées dans le fichier « train.csv ». Ce coefficient nous permet de mesurer la qualité de l’ajustement du modèle aux données observées.



C’est après cela que nous avons utilisé a et b pour prédire les valeurs foncières des logements dans le fichier « test.csv ». Pour chaque arrondissement, nous avons appliqué la formule de prédiction : Valeur foncière prédite = a \* Surface réelle + b



\*

\* \*

En conclusion, nous sommes globalement satisfaits du travail accompli dans le cadre de ce projet. Nous avons consacré des efforts à exploiter les données fournies, ainsi qu’à construire un modèle prédictif des valeurs foncières.

Nous reconnaissons que certaines étapes de notre démarche pourraient avoir été optimisées, notamment en ce qui concerne notre approche initiale lors des premières séances. En effet, nous avons passé beaucoup de temps sur la découverte et la manipulation des données, ce qui aurait pu être évité si nous avions adopté une stratégie plus ciblée dès le début du projet.

Cependant, malgré ces premières difficultés, nous sommes satisfaits des résultats obtenus. Le tri des données nous semble tout à fait convenable et il nous a permis de produire des prédictions cohérentes. De plus, bien que nous ayons choisi un modèle relativement simple, le coefficient de corrélation final semble témoigner de la pertinence de notre approche.