Alioune Diop Chloé Découst

Prédiction des valeurs foncières des logements à Paris

Introduction :

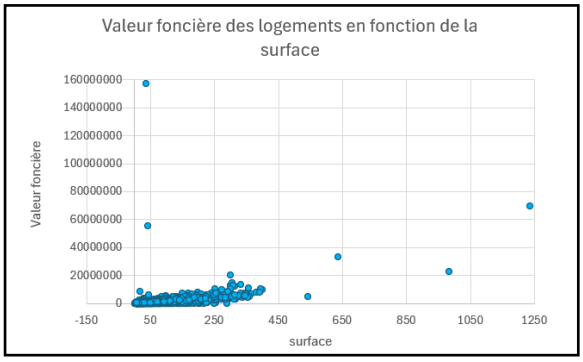
Ce projet a pour but de prédire les prix des logements à Paris. Notre travail consiste à établir le meilleur modèle de régression pour être le plus juste possible sur la prédiction de la valeur foncière. Pour cela, nous avons à notre disposition deux fichiers CSV :

- Un « Train » qui contient le prix de vente et des variables donnant des informations concernant chaque logement qui nous a servis à effectuer nos recherches et nos essais.
- Un « Test » qui contient les mêmes variables que train sans le prix de vente des logements, c'est sur celui-ci qu'il faut prédire les valeurs foncières.

Notre objectif est de rendre un tableau avec l'identifiant des logements concernés et leur valeur foncière prédite pour vérifier qu'elle soit le plus juste possible.

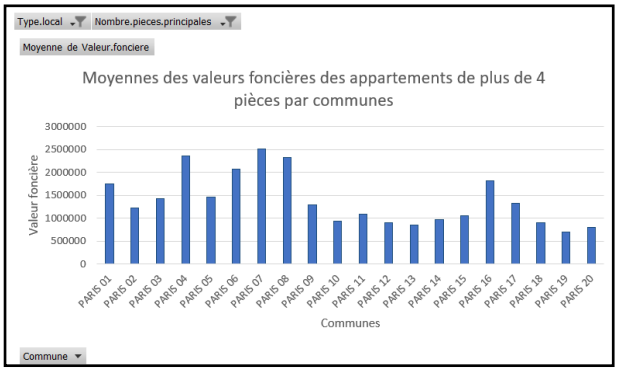
I. Les recherches :

Sur Excel, nous avons utilisé le fichier train pour commencer par établir le nuage de points entre la surface des logements et leur valeur foncière. On a constaté qu'il y avait des valeurs « aberrantes ». On



a donc décidé de supprimer les 10 % des valeurs foncières les plus importantes et les 10 % les plus faibles. Ce qui nous donne plus que 10 240 logements contre 12 796 à la base.

Par la suite, on a décidé d'effectuer des filtres car on peut remarquer qu'il y a une large amplitude que ce soit sur la surface des logements ou leurs valeurs foncières. On a commencé par séparer les maisons et des appartements. Puis on a réparti les apparentements en fonction du nombre de pièces : ceux qui ont moins de 4 pièces et ceux qui ont plus de 4 pièces. Au début, on avait aussi une catégorie d'appartements qui avaient plus de 8 pièces mais nous nous sommes rendu compte, après avoir supprimé les données aberrantes, qu'il n'y avait plus d'appartement de plus de 8 pièces. Ensuite, nous avons calculé la moyenne des valeurs foncières par commune dans chaque catégorie et nous nous sommes aperçus qu'il y a de grandes différences entre les différents arrondissements de Paris. Ce qui nous a menés à faire des sous – catégories entre les logements les plus chers et les moins chers. Cependant,



pour les maisons, nous avons pris la décision de ne pas faire de sous-catégorie parce que nous avons que 6 maisons.

II. Le choix du modèle :

Après avoir établi nos catégories, nous avons testé les 4 modèles pour trouver le plus adapté a notre répartition. Pour cela, nous avons filtré le nombre de pièces et les communes concernés pour effectuer les essais. Nous avons tracé les nuages de points puis la courbe correspondante au modèle que nous testons. On a récupéré l'équation du modèle pour ensuite l'appliquer à nos valeurs pour pouvoir comparer nos prix prédits à ceux du fichier en utilisant la méthode des moindres carrés. On en a conclu que pour chaque catégorie, le modèle avec le plus faible SCR est le modèle linéaire.

III. L'application du modèle :

Sur R, on a commencé par importer nos deux fichiers CSV : train et test. Ensuite, nous avons supprimé nos valeurs aberrantes du fichier train puis on a séparé notre tableau en fonction de nos catégories :

- Appartements plus de 4 pièces les moins chers : appart_plus4pièces_moinschère
- Appartements plus de 4 pièces les plus chers : appart_plus4pièces_pluschère
- Appartements moins de 4 pièces les moins chers : appart_moins4pièces_moinschère
- Appartements moins de 4 pièces les plus chers : appart_moins4pièces_pluschère
- Maison : maison

Par la suite, nous avons cherché le 'a' et le 'b' pour retrouver notre équation correspondante a notre modèle pour chaque catégorie. On a pu estimer les valeurs foncières prédites pour pouvoir les comparer avec celles du fichier train. Par la suite calculer le SCR et comparer avec celui de nos recherches.

Après cela, nous avons réparti les logements présents dans test en fonction des filtres, effectués précédemment sur train, pour appliquer nos modèles. Au final, nous avons 5 data frames différents avec nos valeurs prédites. Il a donc fallu tout les rassembler dans un seul data frame : « prediction » contenant l'identifiant des logements (id) et la valeur foncière prédite (Valeur.fonciere). Pour finir nous avons exporter notre table « prediction » au format CSV.

Conclusion :

Pour finir, nous avons donc décidé d'utiliser le modèle linéaire pour prédire les valeurs foncières en fonction des différentes caractéristiques immobilières telles que la surface, le type de logement et la localisation parce que c'est celui qui nous donnait les meilleures prédictions.

Pour déterminer les arrondissements les plus chers et les moins chers, on aurait pu faire les médianes des moyennes de nos valeurs foncières.