Dauphine | PSL
UNIVERSITÉ PARIS

PROJET DATA MANAGEMENT

M1 ÉCONOMIE & FINANCE

Clara NOCHÉ, Ialgen ALLAL, Sami AMAR, Alice MUNIER

Professeur : M. Daniel Omola

Présentation du sujet

Les trois fichiers mis à disposition dans le cadre de ce projet correspondent aux valeurs foncières des premiers trimestres 2019, 2020 et 2021. Ces fichiers sont, depuis un décret du 28 décembre 2018, disponibles en libre accès. En effet, la publication de ces données répond à l'objectif de transparence des marchés fonciers et immobiliers.
Le jeu de données présenté est produit par la direction générale des finances publiques, qui est une administration de l'État. On peut donc considérer ces données comme fiables. Les données permettent d'avoir des informations sur les transactions immobilières intervenues au cours des 5 dernières années sur le territoire métropolitain et les DOM-TOM, à l'exception de quelques régions (Alsace, Moselle, Mayotte). Les données sont issues des actes notariés et des informations cadastrales.
Les fichiers sont disponibles avec l'extension .csv et contiennent le séparateur suivant : `|`.
Ces données sont très utiles afin d'évaluer le marché immobilier à un moment donné ainsi que son évolution. La valeur foncière permet notamment aux particuliers d'estimer la valeur d'un bien immobilier en consultant ces fichiers.

Les informations des trois fichiers concernent :

- Le prix de vente et la date de transaction d'un bien bâti (appartement, maison) ou non bâti (parcelle, exploitation)
- Le descriptif du bien : type de bien, nombre de pièces, surface en m^2
- La localisation du bien

Nous commencerons, pour notre étude, par considérer l'ensemble des périodes, puis nous ferons un focus sur l'année 2021 (en comparaison à la période 2019-2020).
Nous allons réaliser une analyse exploratoire sur le jeu de données fournis après avoir nettoyé et géré les données non-conformes de la base. Le but sera de résumer les données et d'étudier les corrélations entre les différents éléments d'un même fichier.

Les difficultés du projet résident dans le volume important des données fournies et dans la structure, avec des observations au cours du temps pour un même individu (données de panel).

Les étapes du projet vont être constituées :

- Du nettoyage des données
- De la transformation des données
- D'une analyse exploratoire avec présentations des résultats

Détails des variables gardées pour l'étude :

Date mutation : date de signature du document

Nature mutation : Vente, vente en l'état futur d'achèvement, vente de terrain à bâtir, adjudication, expropriation ou échange
Valeur foncière : Il s'agit du prix ou de l'évaluation déclaré dans le cadre d'une mutation à titre onéreux
Voie : Numéro de la voie
Commune : Libellé de la commune
Code postal
Code département
Type local : maison, appartement, dépendance, local industriel et commercial ou assimilés
Surface réelle bâti : Il s'agit de la somme de la surface réelle du local et des surfaces des dépendances
Surface terrain : Contenance du terrain

Cleansing et Wrangling

La tâche initiale à effectuer lorsque l'on récupère des datasets est la préparation de ces données brutes pour l'exploitation. Cela consiste à nettoyer les données pour qu'elles soient exploitables et les plus pertinentes. On a donc choisi d'organiser nos données de manière à récupérer l'information la plus intéressante. Nous avons rencontré plusieurs types d'erreurs quant à a gestion de ces données brutes.

Pour la construction de notre code nous avons fait le choix de regrouper la partie 'Cleansing' et la partie 'Wrangling'. Certaines étapes de la partie 'Wrangling' nous paraissaient plus judicieux de les insérer avant la fin de la partie 'Cleansing'. La partie 'Wrangling' gère la transformation des données.

Afin de mieux nettoyer nos données brutes, la partie 'Data Selection' a notamment été avancé afin de structurer nos données en gardant les colonnes utiles.

A. Incompleteness

Cette première étape permet de gérer les données manquantes. Les 3 fichiers initiaux comprenaient un total de 43 colonnes. Parmi celles-ci, certaines n'ont que très peu de données et sont donc inexploitables. Afin de récupérer un dataframe simplifié nous allons procéder à un choix quant au traitement de ce manque de données. Dans notre code nous éliminons déjà les colonnes avec plus de 50% de valeurs manquantes. En effet, elles ne permettront pas d'en dégager une analyse car l'échantillon de données serait trop petit. Ainsi, pour l'année 2021 nous avons 3379232 lignes et 43 colonnes. Donc 90% de valeurs manquantes signifie que la colonne avec plus d'environ 3041309 de lignes non renseignées

sera supprimée. Nous affichons un tableau avec le nombre de valeurs manquantes par colonne pour avoir un ordre d'idée.

Ce premier tri retire 21 colonnes pour 2021. Il nous reste 22 colonnes : ['No disposition', 'Date mutation', 'Nature mutation', 'Valeur fonciere', 'No voie', 'Type de voie', 'Code voie', 'Voie', 'Code postal', 'Commune',' Code departement', 'Code commune', 'Section', 'Noplan', '1er lot', 'Nombre de lots', 'Code type local', 'Type local', 'Surface reelle bati', 'Nombre pieces principales', 'Nature culture', 'Surface terrain'].

Puis, étant donné que notre étude se base sur l'analyse des valeurs foncières, il semble que s'il manque une donnée pour cette colonne cela n'est pas intéressant de conserver la ligne. Ainsi, nous choisissons de supprimer l'intégralité des lignes qui ne présentes pas de données pour la colonne 'Valeur fonciere'. Nous ne souhaitons pas essayer de calculer ces valeurs manquantes car nous voulons étudier des valeurs réelles avec le plus d'exactitude possible et remplacer les valeurs manquantes par une valeur calculée biaiserait trop notre étude. Nous supprimons donc 32780 lignes.

Enfin, on supprime toutes les lignes qui n'ont pas de valeurs de surface, que ca soit dans 'Surface reelle bati' ou dans 'Surface terrain'. En effet, la valeur foncière et la donnée de surface sont deux éléments essentiels pour établir notre étude de manière pertinente. Pour cela nous souhaitons nous baser sur des données fiables et réelles qui n'ont pas été trafiquées. Certaines lignes présentent une donnée dans une de ces deux colonnes seulement ('Surface reelle bati' ou dans 'Surface terrain'). Nous créons donc une nouvelle colonne 'Surface' qui prend le maximum entre les valeurs de ces deux colonnes 'Surface reelle bati' et 'Surface terrain' afin que nous puissions avoir pour chaque ligne une donnée de valeur foncière et une donnée de Surface.

B. Duplicates

Nous gérons ici les lignes dupliquées. Nous supprimons donc tous les doublons. 206460 sont lignes supprimées.

C. Data selection

Nous procédons à la sélection des colonnes qui sont essentielles à l'étude que nous porterons sur les valeurs foncières et leurs caractéristiques. La suppression des colonnes non pertinentes dans le cadre de l'analyse des données est donc faite dans cette partie du code. Les 10 colonnes que nous supprimons sont les suivantes : 'No disposition', 'No voie', 'Type de voie', 'Code voie', 'Code commune', 'Section', 'No plan', 'Nombre de lots', 'Code type local','Nombre pieces principales' et 'Nature culture'.
Les 11 colonnes restantes sont : Date mutation, Nature mutation, Valeur fonciere, Voie, Code postal, Commune, Code departement, Type local, Surface reelle bati, Surface terrain et 'Surface' (la colonne que nous avons créée précédemment).

D. Data type conversions

À la suite de l'impotation des données dans le dataframe initial, certaines colonnes n'ont pas le type approprié (string, integer, float). Nous devons donc nous assurer que toutes les valeurs foncières soient des

float et non pas des string par exemple. Nous gérons aussi ce problème de type pour les dates qui doivent suivre le format de date classique jour/mois/année qui est de type datetime. Finalement, pour le code postal nous le convertissons en donnée de type string.

E. Invalidity

Cette étape nous permet de traiter toutes les « fausses » données ou les données invalides. Cela peut être un problème dans l'affichage des données tel qu'un symbole qui ne s'affiche pas correctement ou encore un problème quant à la plausibilité de la donnée tel qu'une valeur foncière négative. Après avoir éliminé les valeurs foncières négatives, nous éliminons les surfaces négatives. Après cette étape aucune ligne n'a été supprimée cela signifie que toutes ces valeurs étaient bien positives.

F. Inconsistency

Cette étape est utile pour vérifier la cohérence des données entre les colonnes. Cela concerne les données entre départements et code postal notamment. Nous vérifions donc pour chaque code postal que celui-ci appartient bien au bon département. Aucune ligne est supprimée ici car il n'y a pas d'incohérence.

G. Aggregation

Plusieurs variables peuvent être agrégées afin d'en ressortir les éléments importants pour l'analyse. Cela permet de consolider des variables similaires. On crée une nouvelle variable dans la colonne 'Prix m2'. Cette colonne permettra de rendre compte du prix au mètre carré pour chaque département par exemple. Ce prix au mètre carré est calculé en divisant la colonne 'valeur fonciere' par la colonne 'Surface'. On a donc pour chaque ligne un Prix m2 car nous nous étions assuré d'avoir pour chaque ligne une valeur foncière et une surface (soit la surface reelle bati soit la surface terrain).

H. Filtration

Nous avons appliqué deux filtres à nos données. Le premier concerne les dates et pour celui-ci on a développé une fonction qui permet de récupérer le mois de chaque date. Cela sera intéressant pour étudier la saisonnalité par exemple en extrayant uniquement les mois sans tenir compte des jours. Le deuxième filtre aussi sous forme de fonction est appliqué sur les départements.

Afin de faciliter notre étude que les surface ('Surface reelle bati' et 'Surface terrain') on a suivi un système d'intervalle. On a créé 10 intervalles pour 'Surface terrain' qui sont stockés dans la nouvelle colonne 'Surface terrain Intervalle'. Ce sont des intervalles de tailles égales. Chaque individu appartient donc à un intervalle en fonction de sa valeur de terrain.

Intervalles	**Nombre d'individus**
(-0.001, 296732.9]	2091381

(296732.9, 593465.8]	184
(593465.8, 890198.7]	28
(890198.7, 1186931.6]	12
(2670596.1, 2967329.0]	5
(1186931.6, 1483664.5]	2
(1483664.5, 1780397.4]	0
(1780397.4, 2077130.3]	0
(2077130.3, 2373863.2]	0
(2373863.2, 2670596.1]	0

Pour 'Surface reelle bat', la taille des intervalles n'est pas égale car les intervalles ont été définis en fonction du nombre d'individus en utilisant la méthode qcut. Ce sont donc des intervalles à taille d'individus quasiment égale que nous avons renommés.

Intervalles	**Nom de l'intervalle**	**Nombre d'individus**
(-0.001, 35.0]	Petit	344235
(35.0, 70.0]	Moyen	346234
(102.0, 237500.0]	Tres grand	330480
(70.0, 102.0]	Grand	327444

Après avoir nettoyé et préparé les données, nous allons les explorer afin d'étudier et comprendre les distributions et les relations entre les données.

L'exploration des données comporte trois étapes majeures :

- *Exploratory data analysis*
- *Feature selection*
- *Feature engineering*

Nous ne ferons pas ici l'étape de *feature selection*
Pour l'analyse exploratoire des données, nous avons réalisé à l'aide de la programmation statistique des graphiques, des diagrammes et autres visualisations pour résumer et observer les données.

Avant de présenter les graphiques réalisés, nous avons émis des hypothèses concernant les relations entre les variables.
Les hypothèses majeures sont les suivantes :

- Il pourrait exister une corrélation linéaire entre le paramètre valeur foncière et prix au m2 étant donné que la valeur foncière correspond à la valeur d'un terrain. Elle est analysée en fonction de son potentiel de construction future.
- Il pourrait aussi y avoir une corrélation positive entre le paramètre valeur foncière et surface réelle bati, pour les mêmes raisons que le point précédent
- Il y a sûrement plus de transactions de maisons et appartements que de dépendances ou locaux industrirels
- On peut s'attendre à augmentation des ventes de maison après la période COVID
- On s'attend à des prix au m2 plus importants dans les zones urbaines et les métropoles par rapport aux zones rurales, et à des prix au m2 plus importants pour les maisons et les appartements par rapport aux locaux industriels
- Le nombre de transactions est sûrement plus important en métropole et en ville plutôt qu'en zone rurale
- Les ventes doivent être majoritaires parmi les *nature mutation,* étant donné que les individus cherchent davantage à acheter des biens aujourd'hui plutôt qu'à les échanger ou les adjudiquer. Aussi, il ne doit pas y avoir d'évolution sur les trois années.

Afin d'étudier les corrélations, nous utiliserons des graphes de type nuages de points. Pour les variables de type string pour lesquelles nous souhaitons étudier la répartition, nous utiliserons des représentations sous forme de camembert et pour des comparaisons zone rurale/zone urbaine, nous utiliserons des tableaux qui permettent de grouper les lignes entre elles selon certains critères. Pour explorer la distribution de nos données, nous utiliserons des graphes de densité

Exploratory Data Analysis

On a choisi d'analyser les colonnes 'Nature mutation' à travers un camembert, car c'est la seule colonne de type string et n'ayant pas trop de valeurs pour qu'un camembert soit illisible.

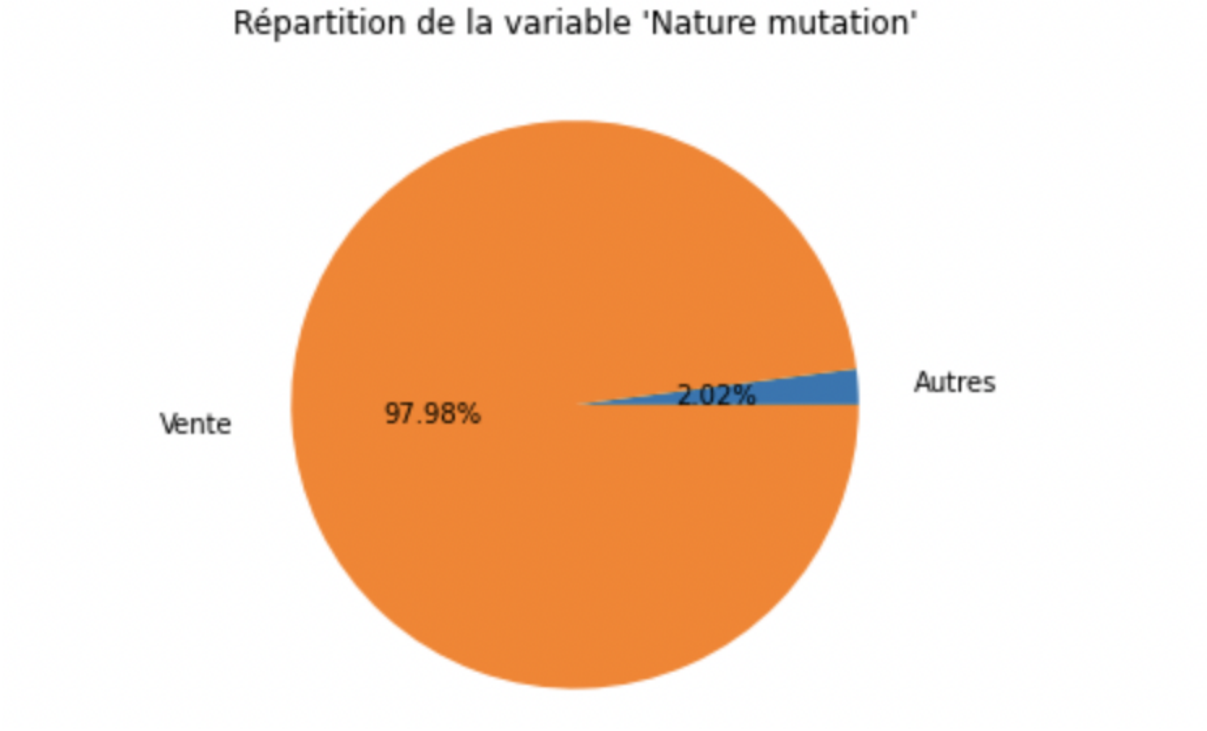
Toutefois, il a quand même été nécessaire de regrouper toutes les catégories ayant une fréquence relative inférieure à 2,5% du total des valeurs de la colonne 'Nombre' du tableau des fréquences.

Pour appliquer cette transformation, nous avons d'abord crée un nouveau data frame qui fait office de tableau de fréquence de chaque colonne. La fonction value_counts permet d'obtenir le tableau de fréquence, tandis que la fonction to_frame permet de transformer ce tableau qui est originairement une série en un data frame. On stocke ensuite l'index dans une colonne s'appelant nom et on recrée un autre index de la même longueur que le data frame avec range.

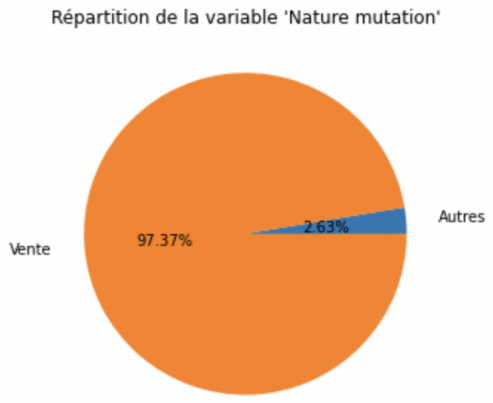
Grace à l'indexation booléenne et loc, on change toutes les cases de la colonne 'Nom' ayant une fréquence inférieure à 2,5% à 'Autres'. Ensuite avec groupby, on fusionne toutes les cases 'Autres' pour en obtenir qu'une seule. On n'oublie pas de réinitialiser l'index avec la fonction reset_index pour éviter les erreurs.

On rend le processus itérable avec une boucle for qui parcours x dans une liste des colonnes concernées, et en définissant le tableau de fréquence par rapport à x.

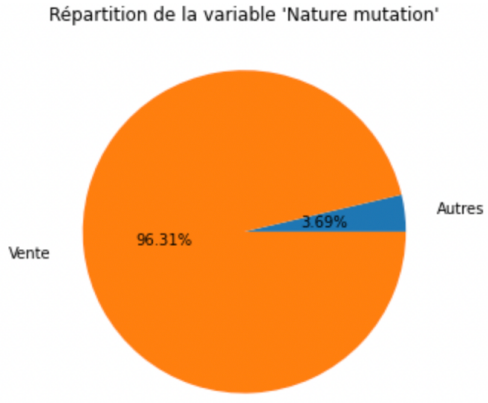
2021



2020

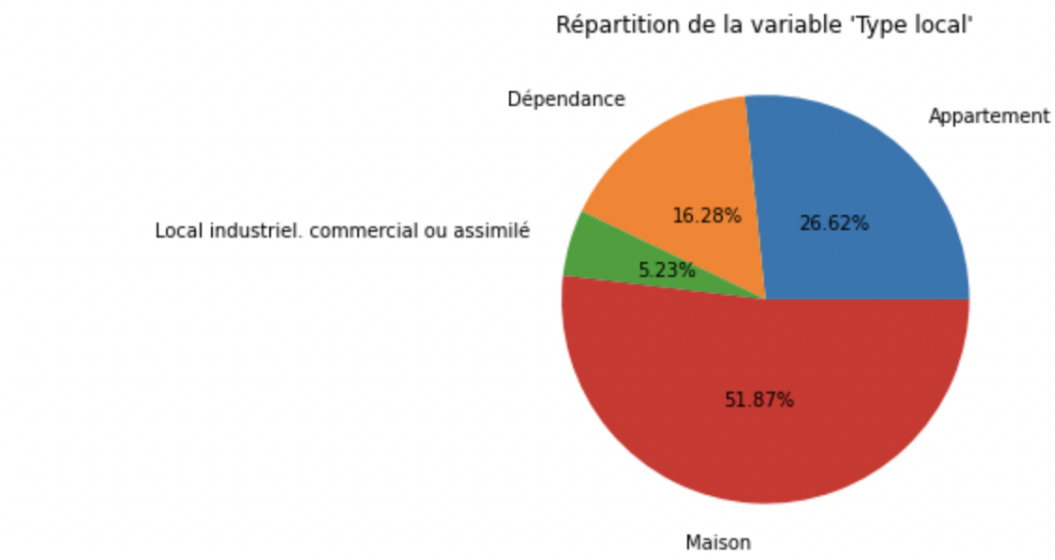


2019

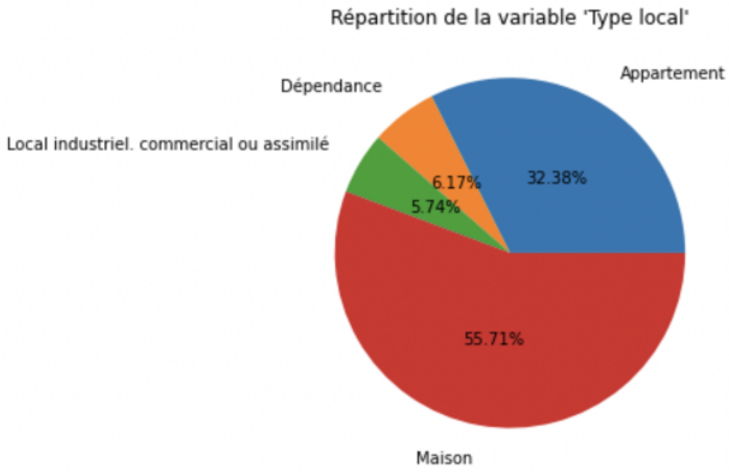


On remarque que la très grande majorité des mutations se font par vente, et que les mutations du type : échange, vente terrain à bâtir, adjudication et expropriation comptent pour seulement un peu plus de 2% des transactions. Notre hypothèse est donc vérifiée ici, les ventes sont largement majoritaires, et les pourcentages sont les mêmes pour les trois années.

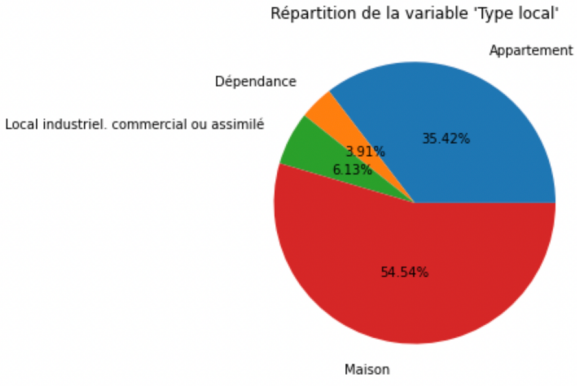
2021



2020

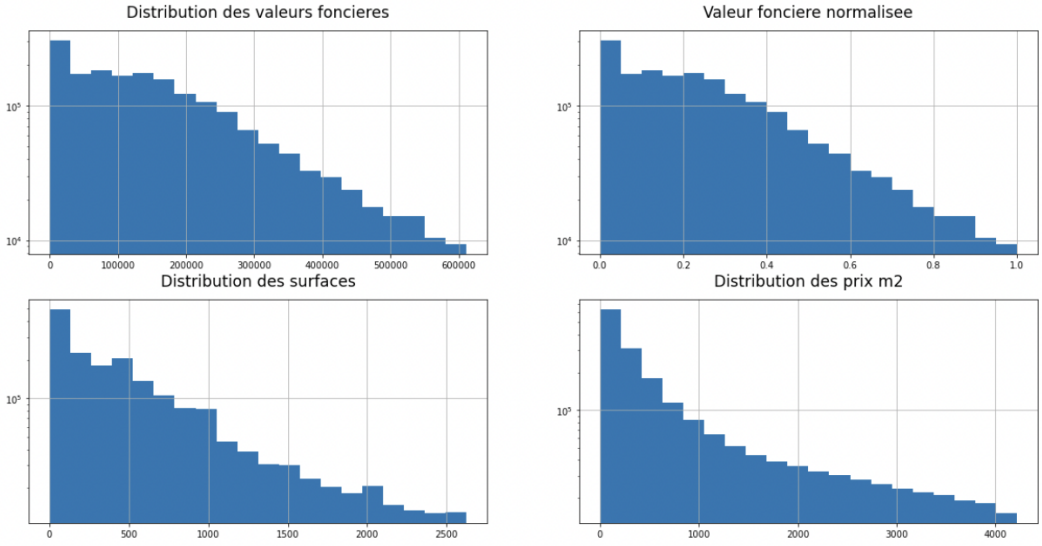


2019



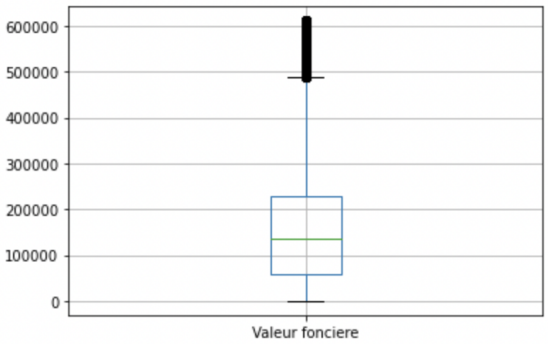
On remarque ici que les proportions du type de local sont assez constantes sur les trois années étudiées. Notre hypothèse selon laquelle davantage de maisons ont été vendues en 2020 pendant la pandémie de COVID est vraie mais pas vraiment significative étant donné qu'on passe de 54,5% de maisons vendues en 2019 à 55,7% en 2020.

2021



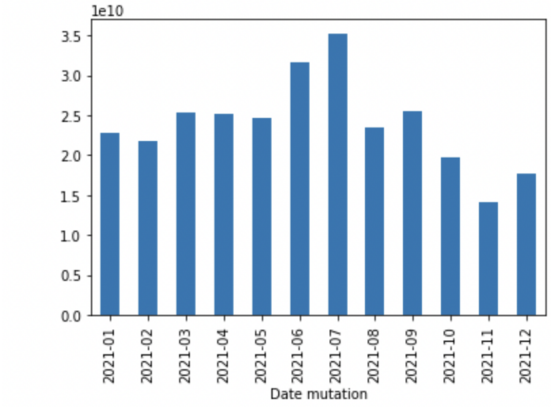
Les quatre graphes des distributions présentés ci-dessus sont quasiment identiques pour 2019, 2020 et 2021.
La distribution de ces quatre graphiques montre que les variables suivent une loi exponentielle. Les échelles des graphiques sont en log. Cette transformation en log permet de mieux interpréter toutes les valeurs. En effet, la répartition est plus étalée, on peut donc se rendre compte de l'importance du nombre de valeurs proche de 0. Cependant, il ne faut pas négliger les valeurs qui s'en écartent. Les valeurs des surfaces sont largement concentrées entre 0 et 1000. Pour les valeurs foncières il y en a autant entre 0 et 200000.

2021

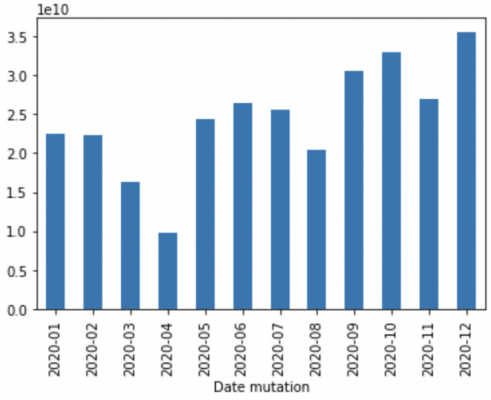


Les boîtes à moustaches sont quasiment identiques pour les trois années, c'est pourquoi nous ne mettons que celle-ci.
La valeur maximale pour la valeur foncière est de 500000, la valeur minimale de 0. Le quartile inférieur est de 70000 environ, le quartile supérieur de 220000 environ et la moyenne de 140000. On a des valeurs extrêmes et atypiques entre 500000 et 600000. La dispersion est ici importante étant donné la longueur de la boîte à moustache.

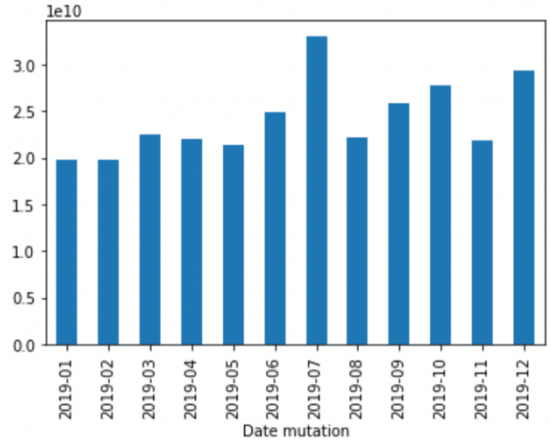
2021



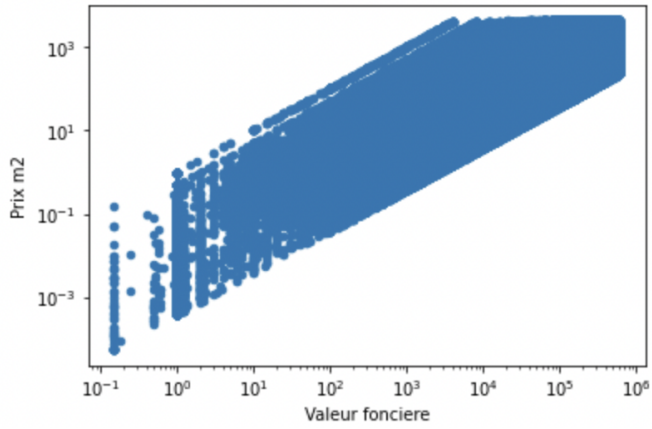
2020



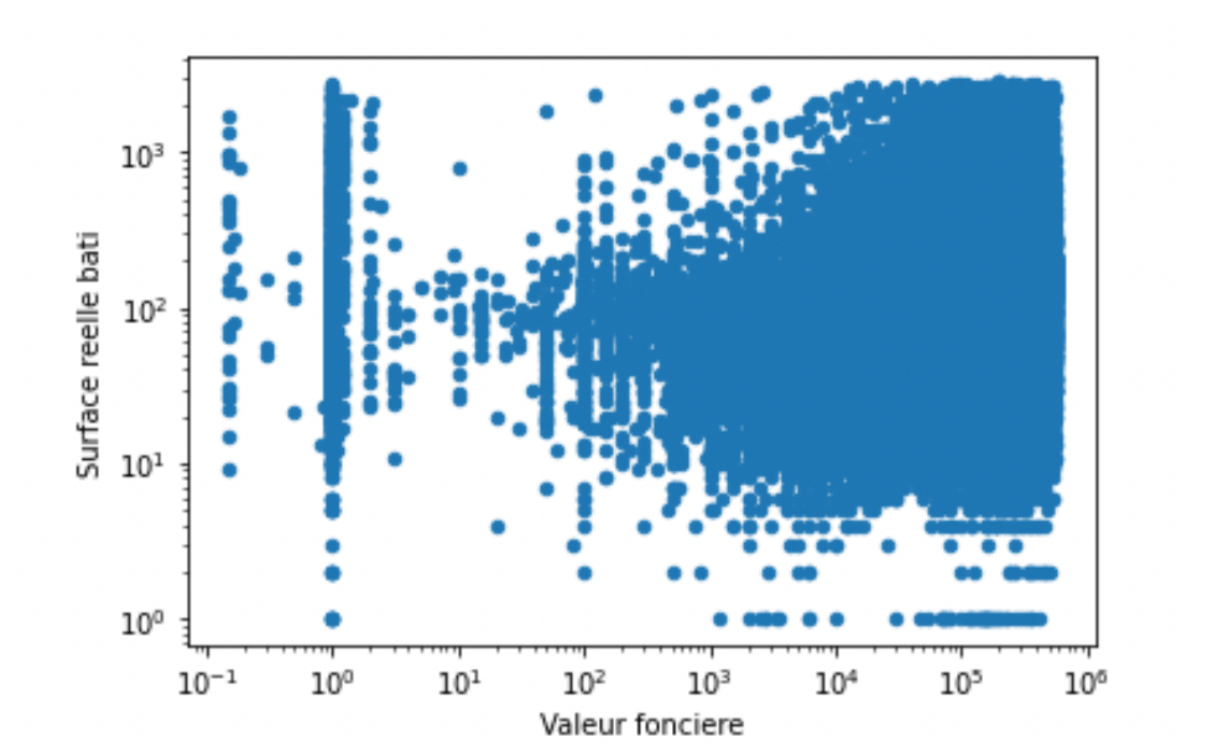
2019



2021



Le graphe en nuage de points est le même pour 2020 et 2019. On remarque comme prévu qu'il y a bien une corrélation positive et linéaire entre le prix au m2 et la valeur foncière, ce qui est tout à fait logique étant donné que la valeur foncière correspond à la valeur d'un terrain.



On constate une très grande répartition hétérogène. Cela signifie que chaque surface réelle bati peut être attribué à une valeur foncière donc à un prix. Cela est cohérent car en fonction des départements on observe de larges différences quant au prix des mètres carrés.
Cependant, il y a un alignement des points sur les valeurs foncières de 10 et de 10^{-1}. Ceci peut être plus difficile à interpréter mais cela peut être dû au manque de données pour les valeurs foncières.

2021

Commune	Prix m2
ROUVROIS SUR OTHAIN	0.050725
CESSE	0.763359
FRAIN	0.934579
CHEVROCHES	1.126245
SAINT-GEORGES-DES-AGOUTS	1.234568
...	...
BAUME (LA)	3644.871795
LYON 7EME	3692.592593
PIERREFITTE-EN-AUGE	3808.703704
BOURGEAUVILLE	3845.930233
MERIA	3933.333333

2020

Commune	Prix m2
BAUDRECOURT	0.000725
PARIS 17	0.002500
FERNEY-VOLTAIRE	0.003623
VAUDONCOURT	0.014286
DARDENAC	1.596694
...	...
NEUILLY-SUR-SEINE	3873.239437
PARIS 20	3890.909091
VERO	3993.243243
CENTURI	4125.000000
VIGNEC	4210.937500

2019

Commune	Prix m2
LEVALLOIS-PERRET	0.001355
BEURVILLE	0.002519
CLICHY	0.002564
LES EPARGES	0.004167
WARLUZEL	0.006180
...	...
LYON 1ER	3800.000000
PARIS 20	3888.888889
BOULOGNE-BILLANCOURT	3928.999129
ABRIES	3932.478261
FOURQUEUX	3941.525424

On voit ici que les métropoles et villes chères de banlieue figurent dans le top 5 des villes avec les prix au m2 les plus importants. Le top 5 des villes avec les prix au m2 les plus faibles sont cependant plus surprenant. Cette incohérence est la cause d'un manque de données. Aussi en 2021, on ne retrouve pas Paris ou une ville de sa banlieue dans le top 5 des plus hauts prix au m2 comparé aux années précédentes.

2021

Code departement	Prix m2
23	259.328095
58	287.836653
70	291.076631
36	292.678725
52	301.601045
...	...
78	1083.863190
95	1171.887550
93	1190.821566
94	1422.866462
92	1937.557030

2020

Code departement	Prix m2
23	230.791502
70	283.647397
36	292.906515
58	295.317836
88	296.803772
...	...
13	1083.983074
95	1109.456205
93	1166.809427
94	1354.663669
92	1933.463756

2019

Code departement	Prix m2
23	224.736581
70	254.098755
58	262.580306
36	282.685768
52	286.548774
...	...
95	1063.173858
13	1083.685304
93	1110.071205
94	1313.329530
92	1991.940272

Les départements qui, sur les trois années, comportent des prix au m2 les plus hauts sont tous des départements de banlieue de Paris (95, 94, 93, 92) sauf le 13. Ceci paraît très cohérent. En effet, les banlieues de Paris se sont gentrifiées ces dernières années, avec des personnes cherchant de plus en plus à acheter des biens avec un terrain, donc la demande pour des biens en banlieue de Paris a augmenté faisant augmenter les prix au m2. Concernant les 5 départements avec les prix au m2 les plus bas on retrouve que des départements ruraux, comme évoqué dans nos hypothèses.

2021

Type local	Prix m2
Maison	580.248029
Local industriel. commercial ou assimilé	672.266625
Appartement	1054.636962

2020

Type local	Prix m2
Maison	538.936499
Local industriel. commercial ou assimilé	642.181702
Appartement	1015.842619

2019

Type local	Prix m2
Maison	538.936499
Local industriel. commercial ou assimilé	642.181702
Appartement	1015.842619

Les chiffres sont assez stables sur les trois années pour ce tableau. Une maison est généralement accompagnée d'un terrain qui vaut moins cher et il y a beaucoup plus de maisons dans les endroits ruraux. En revanche, les appartements sont plus souvent situés en ville donc dans des endroits avec un prix au mètre carré plus cher. Ces chiffres sont cohérents avec la réalité des prix de l'immobilier.

Pour conclure, nous pouvons dire que nos données semblent cohérentes et la majorité de nos hypothèses sont confirmées. Cependant, certaines données importantes sont manquantes comme la surface de biens vendus avec un prix élevé, ce qui biaise les résultats et les distributions. Ainsi, une base de données plus complète permettrait une meilleure analyse des transactions immobilières ces trois dernières années. Nous avons aussi procédé à différents choix concernant le traitement de nos données. Si ceux-ci avaient été différents nous aurions pu obtenir des données différentes. Les choix que nous avons effectués sont notamment de supprimer des lignes lorsque des valeurs extrêmes ont été repérées. Nous aurions pu leur attribuer des valeurs en se basant sur une moyenne des prix en fonction des départements, ce qui aurait permis une analyse plus complète. La sélection initiale des colonnes peut aussi s'avérer être un autre biais. Le nettoyage des données brutes a été fait de manière arbitraire en privilégiant une étude entre valeur foncière et surface. Il serait intéressant à présent de se consacrer à une étude qui intègre des dates antérieures. En effet, l'évolution du marché immobilier a énormément changé ces dernières années. On pourra notamment remarquer un point de retournement clé avec la crise de 2008. Des comparaisons intéressantes sur l'impact de cette crise pourront être faites.