M1 AEEF - 2022/23
Joudy Benkaddour
Gabrielle Morin

Data Management
Analyse de l'impact de la crise sanitaire sur le marché de l'immobilier français

« Je déclare sur l'honneur que ce mémoire a été rédigé de ma main, sans aide extérieure non autorisée, qu'il n'a pas été présenté auparavant pour évaluation et qu'il n'a jamais été publié, dans sa totalité ou en partie. Toutes parties, groupes de mots ou idées, aussi limités soient-ils, y compris des tableaux, graphiques, cartes etc. qui sont empruntés ou qui font référence à d'autres sources bibliographiques sont présentés comme tels, sans exception aucune. »

TABLE DES MATIERES

INTRODUCTION

La crise sanitaire de la Covid-19 ont profondément impacté nos habitudes de consommation. Ainsi, les confinements successifs ont conduit à une chute de profit dans les secteurs du service (tourisme, restauration, loisirs) tandis que la vente en ligne ou les plateformes de vidéos à la demande ont réalisé d'important profit.
Ces changements de consommation ont été associé à un profond bouleversement des marchés financiers avec des chute de valorisation très importantes (le CAC 40 a chuté de 33% entre le 14 février et le 20 mars).
Cette crise a donc conduit les gouvernements à soutenir massivement l'économie à l'aide de primes pour les consommateurs et de soutiens aux entreprises afin d'éviter leurs faillites.

Aux vues de ces changements, nous nous intéressons au marché de l'immobilier. En effet, celui-ci est réputé pour être stable et une valeur refuge dans le temps ; c'est-à-dire qu'investir dans l'immobilier permet de garder son capital intact quel que soit les états de marché. Or, la crise de la Covid-19 a été un grand bouleversement pour la valorisation des actifs financiers malgré le soutien des gouvernements
Le but de ce projet est donc d'évaluer l'impact de la crise sanitaire sur les prix du marché immobilier français.

Après avoir importé et traité notre base de données, nous étudierons la relation entre nos variables. Ensuite, nous étudierons l'impact de la crise sur le marché immobilier dans son ensemble avant d'affiner notre analyse par région puis par département. Enfin, nous tenterons de créer un modèle prédictif des prix de l'immobilier en France.

1. IMPORTATION DES DONNEES

Dans ce projet, nous utilisons des données provenant du Ministère de l'économie, de la finance et de la relance. Nous importons les fichiers de demandes de valeurs foncières du second semestre de 2017 au premier semestre de 2022.
Ces jeux de fichiers permettent de connaître les transactions immobilières effectuées au cours des cinq dernières années sur le territoire français (métropole et DOM-TOM), à l'exception de l'Alsace, de la Moselle, et de Mayotte. Les données sont issues des actes notariés et des informations cadastrales.
Ces fichiers sont mis à jour de manière semestrielle en avril et octobre. Les données disponibles ont été mises à jour le 05 avril 2023. Nous avons donc ajouté le second semestre 2022 à notre jeu de données (à noter que le second semestre de 2017 n'est plus disponible sur le site internet mais que nous l'avons gardé dans notre jeu de données).

Pour ce faire, nous importons les bases de données dans différentes variables. Etant donné qu'elles sont toutes extrêmement lourds (43 colonnes, plus de 1,5 millions de lignes chacune), nous faisons le choix de les retraiter avant de concaténer les données en une seule base de données.
Nous avons également réalisé une fonction permettant de voir les différentes caractéristiques de nos bases de données : statistiques descriptives, 5 premières lignes, dimensions et taille de la base.
Cette fonction nous donne une première idée de comment retraiter notre base de données de manière efficace. Ainsi, nous pouvons voir que les colonnes de nos bases de données semblent être identiques, nous le confirmerons plus tard à l'aide d'une fonction, mais l'on sait déjà qu'il sera possible de concaténer nos bases de données en une seule. De plus, nous pouvons voir que toutes nos colonnes ne se sont pas importées dans le bon format (dates, valeur foncière, et code postal), problème qui sera adressé dans la partie 2.2.2.

Cette première partie a été la plus complexe. En effet, la quantité de données à traiter est très importante, ce qui a conduit à de multiples erreurs sur nos ordinateurs respectifs (manque de mémoire, de ram, et dépassement de performance). Nous avons donc dû découper nos bases de données, et les importer une à une depuis nos ordinateurs (et non l'url du site) afin que cela fonctionne. Tout au long de ce projet, nous avons également dû faire attention à la quantité de mémoire qui était utilisée par le processeur pour éviter que le code ne s'arrête brutalement.

Une fois les bases de données importées, nous les retraitons afin qu'elles soient dans un format approprié pour l'analyse des données.

2. TRAITEMENT DE LA BASE DE DONNEES

Maintenant que nous avons importé nos données, nous devons préparer notre base de données afin de pouvoir l'utiliser.
Nous retraitons donc certaines colonnes, le type de nos valeurs, les valeurs nulles, et ajoutons les colonnes nécessaires à notre analyse.

2.1. Retraitement des colonnes

Nous remarquons qu'en 2022, le nom de la première colonne du fichier source a changé et est devenus *Identifiant de document* (auparavant *Code service CH*). Or, le document descriptif des fichiers nous indique que le contenu des deux colonnes est identique. Nous changeons donc le nom de la colonne en *Code service CH* pour pouvoir concaténer les bases de données. A noter que c'est un changement temporaire car les deux colonnes (*Identifiant de document* et *Code service CH*) ne contiennent aucunes données.

Les bases de données que nous avons importées, et que nous avons concaténées, sont extrêmement lourdes et volumineuses, nous supprimons donc les colonnes inutiles afin d'accroître la rapidité d'exécution de notre code.

Certaines colonnes des fichiers sources sont vides, nous les supprimons donc car non-pertinents pour notre analyse :

- Le code service CH ou Identifiant de document ;
- La référence document ;
- Les cinq colonnes contenant les articles du code général des impôts (CGI) ;
- Identifiant local.

De plus, certaines colonnes ne nous seront pas utiles dans notre analyse, nous les supprimons également :

<u>Le numéro de disposition :</u>
Une disposition constitue une composante d'analyse juridique. L'analyse n'ayant aucun trait juridique, nous avons préféré écarter cette donnée.

<u>N° de voie ; B/T/Q ; Type de voie ; Code de voie ; Voie :</u>
Notre analyse se concentre sur le marché de l'immobilier au niveau national, régional et départemental, ces données sont donc superflues.

<u>Code commune :</u>
Nous effectuons une analyse au niveau régional et départemental, nous n'avons donc pas besoin d'un tel niveau de détail. Dans le cas où nous déciderions de pousser l'analyse, nous avons gardé le nom de la commune en toutes lettres.

Préfix de section ; section :
Identifiant utilisé pour différencier les immeubles pour le cas de certains quartiers ou des communes absorbées. De la même manière que pour l'adresse ou le code de la commune, nous n'avons pas besoin d'un tel niveau de détail.

No plan ; No volume :
Surface situées en dessous ou en dessous du bien vendus. Nous ne considérons pas cela comme un facteur influençant la valeur foncière d'un bien vendus.

X^{ème} lot ; Surface Carrez du X^{ème} lot :
Nous ne nous intéressons pas au nombre de lots dans chaque bien ni au détail de la surface Carrez desdits lots. En effet, seul la surface des cinq premiers lots apparaît dans ce document. Nous ne considérons donc pas ces colonnes comme assez complètes pour les utiliser.
Pour mesurer la surface d'un bien nous préférons utiliser la surface réelle bâtie (incluant les espaces considérés comme non-habitables) ainsi que la surface du terrain, si le bien vendus présente un terrain.

Nombre de lots :
Nombre de lots de copropriétés dans le bien vendu. Nous avons jugé ce niveau de détail trop important pour notre analyse.

Code type local :
Nous avons préféré garder le type de local en toute lettres pour des questions de lisibilité. Garder cette colonne aurait un effet redondant.

Nature culture ; nature culture spéciale :
Nature des cultures présentes sur les terrains (présence de forêts, ou étangs par exemple). Nous effectuons une analyse générale du marché de l'immobilier français, ce niveau de détail est donc trop important.

2.2. Retraitement des variables

2.2.1. VALEURS NULLES

Certaines colonnes présentent des valeurs nulles qu'il convient de retraiter. Nous avons donc effectué les retraitements suivants :

Valeur foncière :
Les biens sans valeur foncière sont des donations. Ils ont une valeur mais celle-ci n'est pas reflétée dans la valeur d'échange. Nous transformons donc les valeurs nulles en 0.

Code postal :
La colonne *Code départemental* ne présente aucune valeur nulle ; l'absence de code postaux pour toute une série de biens n'est donc pas dérangeante. En effet, nous effectuons une analyse régionale puis départementale ; le code départemental nous suffit pour obtenir ces informations. Les valeurs nulles sont donc remplacées par un tiret.

Type local :
Les cellules ne présentant pas de valeurs sont les biens sans surface réelle bâtie mais possédant une surface terrain. Cela nous a permis de déterminer que ces biens sont tout simplement des terrains. Nous avons donc transformé les valeurs nulles de cette colonne en « terrain ».

Surface réelle bâtie :
Les cellules ne présentant pas de valeurs sont les transactions de terrains sans constructions (prairies, champs ou forêts par exemple). Les valeurs nulles sont donc transformées en 0.

Nombre de pièces principales :
Les biens sans pièces principales étant des terrains, nous avons transformé les valeurs nulles de cette colonne en 0.

Surface terrain :
De la même manière que pour la surface réelle bâtie, les valeurs nulles sont transformées en 0. En effet, certains bien vendus n'ont pas de terrain associé à la surface bâtie. C'est par exemple le cas d'appartements, bureaux, ou maisons de ville.

2.2.2. TYPE DE DONNEES

Lors de l'importation de nos données, certaines variables n'ont pas été importées dans le bon type, ce qui pose problème lors du traitement et de l'analyse.

Date :
Les dates (première colonne de la base de données) ont été importées en tant que chaînes de caractère. Or, nous effectuerons des graphiques en fonction du temps lors de l'analyse de l'impact de la crise sanitaire. Nous avons donc transformé la chaîne de caractère en date.

Valeur foncière :
La valeur foncière était importée en tant que chaîne de caractère. Or, c'est un nombre (de type *float*) sur lequel nous allons effectuer des analyses.
Pour effectuer la conversion, nous devons traiter la différence de convention entre les chiffres français et anglais. Notre fichier source étant français, les décimales sont indiquées par une virgule tandis que python est un logiciel anglais. La décimale est donc indiquée par un point. Nous remplaçons donc les virgules de la colonne par des points avant de convertir les chaînes de caractère en *float*.

Code postal :
Le code postal était importé en tant que nombre (avec une décimale) alors que nous le traiterons en tant que chaîne de caractère dans notre analyse.
Nous avons donc supprimé la décimale et convertit le code postal en chaîne de caractère

2.3. Retraitement des valeurs

2.3.1. TYPE DE TRANSACTION

Nous faisons le choix de garder uniquement les transactions étant des ventes simples. Nous supprimons donc les lignes des transactions suivantes :

- Adjudications ;
- Echanges ;
- Expropriation.

Nous considérons en effet que ces biens ne sont pas représentatifs du marché de l'immobilier français, car les transactions n'opèrent pas de manière régulière.

Nous décidons également d'analyser uniquement le marché de l'immobilier des biens finis. De ce fait, nous supprimons toutes les lignes de ventes de terrains à bâtir ou les ventes en l'état futur d'achèvement.

2.3.2. RETRAITEMENT DES VALEURS EXTREMES

Lors de l'importation des données, nous avons pu regarder les statistiques annuelles de l'échange de biens. Les bases de données présentent des valeurs extrêmes qui ne reflètent pas la réalité du marché immobilier français et fausseraient les estimations que nous effectuerons.

Nous avons donc décidé de supprimer toutes les lignes dont la valeur foncière est inférieure à 500 euros, soit 287 823 lignes. De cette manière, nous retraitons les donations, les biens échangés pour une valeur symboliques, ainsi que les biens en très mauvais états.

De la même manière, nous supprimons les 99ème quantile des valeurs les plus élevées pour retraiter les biens exceptionnels vendus extrêmement chers : hôtels particuliers ou villas de luxe, immeubles ou immenses zones bâtimentaires. Nous supprimons donc toutes les biens ayant une valeur foncière supérieure à 15 000 000 €, soit 183 201 lignes.

Le retraitement des valeurs extrêmes nous a permis de passer de 18 617 295 lignes à 17 946 450 lignes, soit une réduction de 3,6%.

2.4. Ajout de colonnes

2.4.1. REGIONS ET DEPARTEMENTS

Nous allons étudier l'impact de la crise sanitaire sur le marché de l'immobilier au niveau des régions françaises. Or, le document nous fournit uniquement le département (colonne *Code département*) dans laquelle la transaction a eu lieu.

Nous créons donc une nouvelle colonne, à la fin de la base de données, contenant la région où la transaction a eu lieux en utilisant une fonction de mapping faisant correspondre le code du département au nom de la région à l'aide d'un dictionnaire créé au préalable.

De la même manière, nous effectuons un mapping pour obtenir le nom du département où le bien a été vendu en toute lettre et non un simple code comme dans le fichier d'origine.

2.4.2. LA SURFACE TOTALE

Lors de notre analyse des variables en partie 3, nous nous sommes rendu compte que les valeurs nulles de la surface réelle bâtie et de la surface terrain faussaient la relation entre nos variables. En effet, la valeur foncière d'un bien n'est pas conditionnelle à la taille de son terrain ni à la surface bâtie. Ainsi, nous avons décidé d'étudier la valeur foncière en fonction de la surface du bien, qu'elle soit bâtie, ou qu'elle soit un terrain.
Pour ce faire, nous avons créé une nouvelle colonne dans notre base de données dans laquelle nous avons sommé les deux colonnes de surface (celle réelle bâti, et celle du terrain).

2.4.3. MOIS & ANNEE

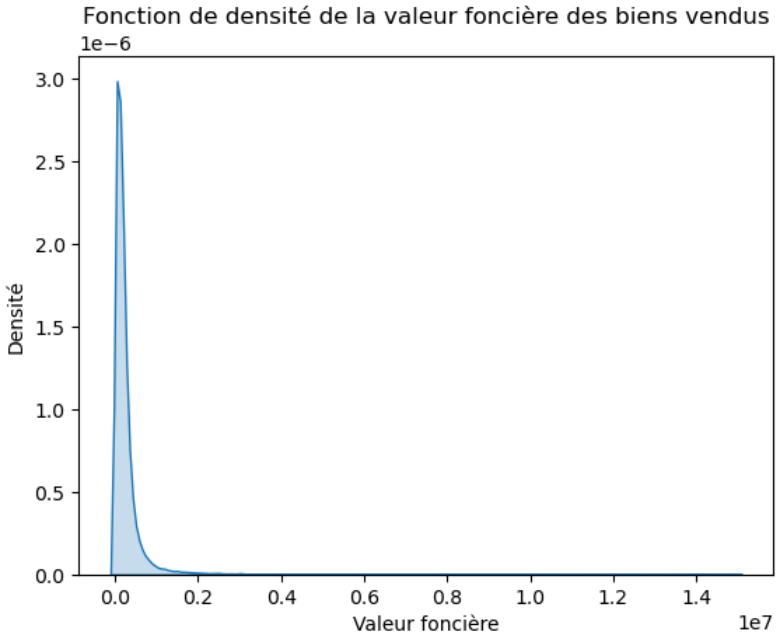
Enfin, nous ajoutons deux colonnes contenant uniquement le mois (pour la première), et l'année (pour la seconde) de la transaction. Cette colonne nous permettra de regrouper nos données lors de notre analyse dans les parties 3 et 4.
Elles nous permettent également un rendu plus lisible lorsque l'on effectue des graphiques dans les parties 3 et 4.

3. ANALYSE DES VARIABLES SUSCEPTIBLES D'INFLUENCER LES PRIX DE L'IMMOBILIER

3.1. Valeur foncière

La valeur foncière d'un bien correspond à son prix. Nous cherchons donc à visualiser cette donnée pour voir ce que nous pouvons en apprendre.

Nous commençons par visualiser la fonction de densité de la valeur foncière.



Nous pouvons voir que la grande majorité des biens ont une valeur inférieure à 2,5 millions d'euros. Malgré notre retraitement des valeurs extrêmes, certains biens présentent toujours une valeur de plus de 14 millions. Ces observations graphiques sont confirmées par l'analyse des statistiques descriptives de la variable :

	Valeur
Nombre de valeurs	17 946 450
Moyenne	344 337 €
Minimum	500 €
1er quantile	66 000 €
Médiane	150 000 €
3eme quantile	279 000 €
Maximum	14 994 550 €
Skewness	8,44
Kurtosis	84,96

Nous pouvons ainsi voir que les trois quarts des biens ont une valeur foncière inférieure à 280 000 euros ; notre jeu de données présente toujours des valeurs extrêmes.

Nous notons également que la moyenne est de deux fois supérieure à la médiane, la valeur foncière de ces valeurs extrêmes tire donc la moyenne vers le haut. Celle-ci n'est donc pas représentative du jeu de données.
Enfin, il est peu surprenant que notre jeu de données ne suive pas une loi normale ; notamment à cause du coefficient d'acuité qui mesure la taille des queux de distribution et qui est très élevé.

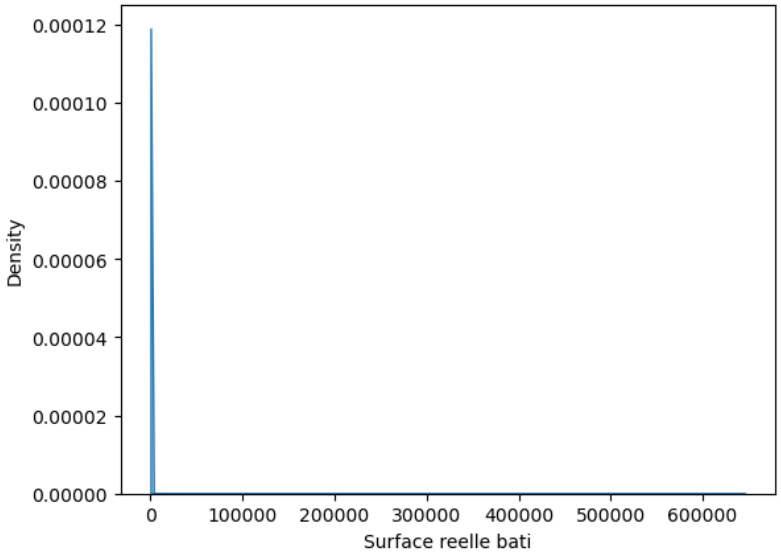
3.2. Surface du bien

La surface d'un bien est l'une de ses caractéristiques les plus importantes. Nous nous intéressons ici à la relation entre valeur foncière d'un bien et la surface, bâtie ou du terrain.

3.2.1. SURFACE REELLE BATIE

Nous commençons par analyser la fonction de densité et les statistiques de la surface réelle bâtie.

<u>Fonction de densité :</u>



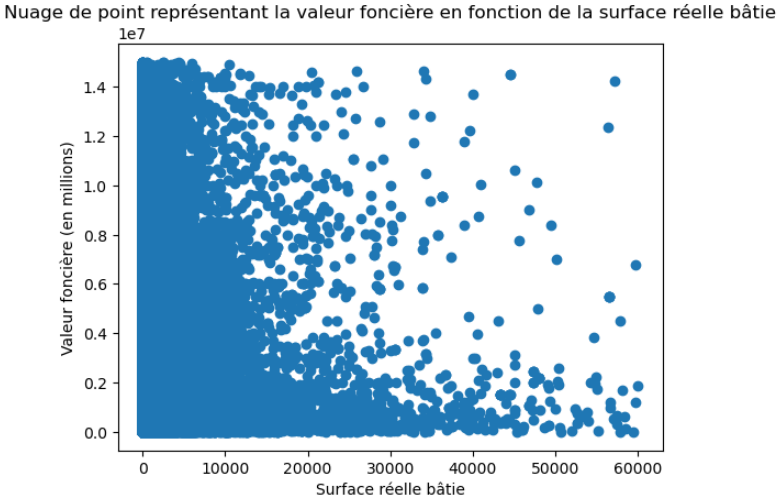
<u>Statistiques descriptives :</u>

	Valeur
Moyenne	46 m²
Minimum	0 m²
1er quantile	0 m²
Médiane	0 m²
3eme quantile	65 m²
Maximum	646 230 m²

Nous pouvons voir que plus de la moitié des biens n'ont pas de surface réelle bâtie, ce qui laisse penser que cet indicateur n'est pas adéquat car il prend uniquement les surfaces construites en compte.

Nous remarquons également la présence de valeurs extrêmes avec un bâtiment ayant une surface de plus de 600 000 m². Il pourrait s'agir d'un hangar, d'une tour de bureaux ou d'un lot regroupant plusieurs biens. Cependant, il s'agit d'une transaction irrégulière que nous retraitons. Ainsi, nous supprimons tous les biens dont la surface réelle bâtie est supérieure à 60 000 m².

Nous visualisons la relation entre valeur foncière et surface réelle bâtie.

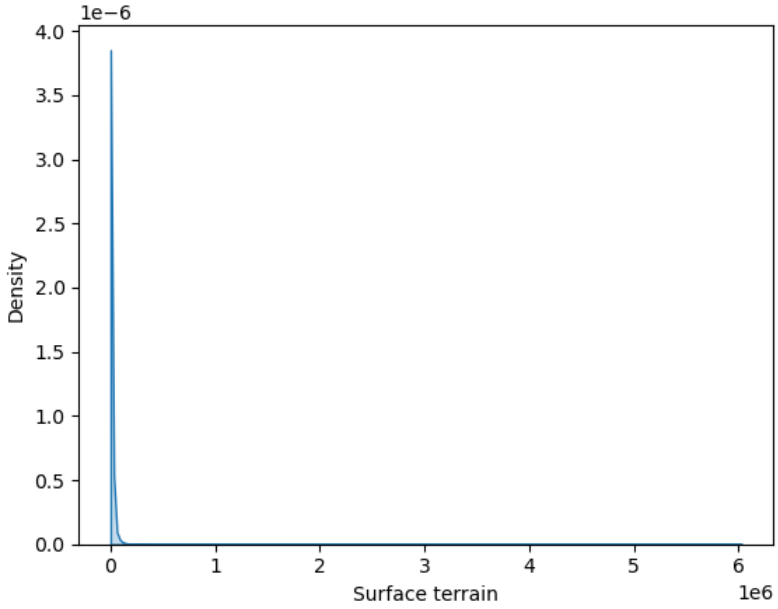


La valeur foncière d'un bien ne peut pas être modélisé uniquement par la surface réelle bâtie. En effet, plus de la moitié des biens vendus ne sont pas construits.

3.2.2. SURFACE TERRAIN

Nous nous intéressons maintenant à la surface du terrain des bien vendus.

<u>Fonction de densité :</u>

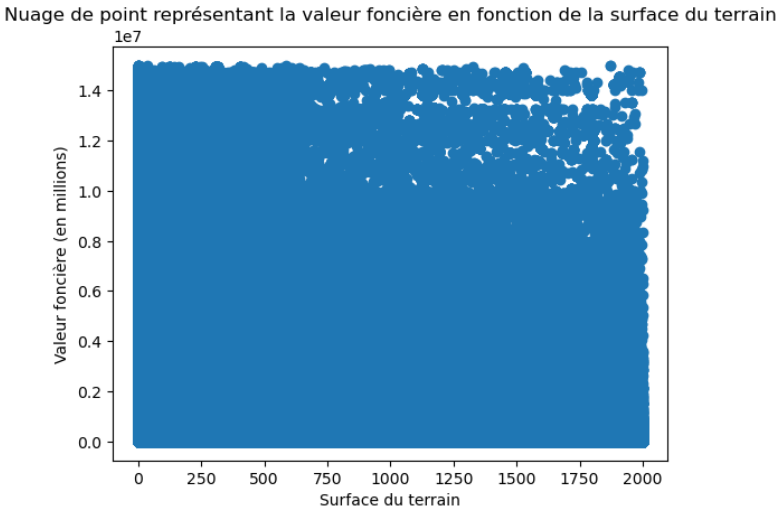


Statistiques descriptives :

	Valeur
Moyenne	2 094 m²
Minimum	0 m²
1er quantile	0 m²
Médiane	317 m²
3eme quantile	1 010 m²
Maximum	6 032 439 m²

Nous sommes confrontés au même problème que pour la surface réelle bâtie, une partie de notre échantillon ne possède pas de terrain dans son bien.
Nous remarquons encore une fois la présence de biens avec un terrain extrêmement grand, ce qui n'est pas représentatif du marché de l'immobilier français. Nous supprimons donc tous les biens ayant un terrain de plus de 2 000 m².

Nous visualisons la relation entre valeur foncière et surface du terrain.

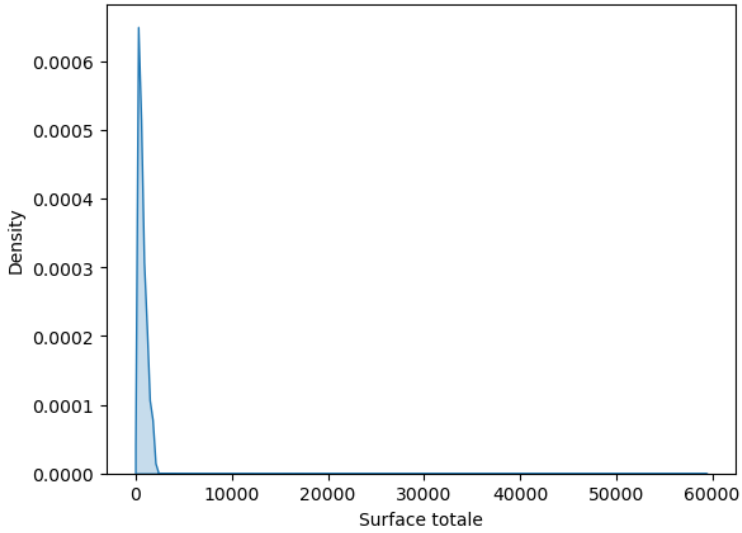


Ce graphique est illisible, la seule observation que l'on pourrait fournir est que pour chaque surface de terrain correspond une valeur possible.

3.2.3. SURFACE TOTALE

Etant donné qu'analyser la surface réelle du bien et la surface du terrain séparément n'a pas donné de résultats concluants, nous analysons la surface totale des biens (soit la somme entre surface réelle bâtie et surface terrain).

Fonction de densité :

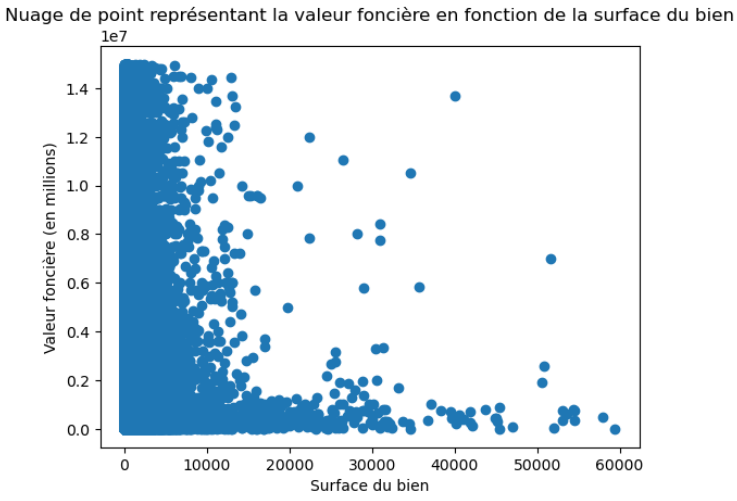


Statistiques descriptives :

	Valeur
Moyenne	405 m²
Minimum	0 m²
1er quantile	39 m²
Médiane	207 m²
3eme quantile	637 m²
Maximum	59 394 m²

Nous pouvons voir que notre échantillon est plus cohérent que lorsque l'on étudie les surfaces bâties et terrains séparément.
Nous notons cependant qu'une partie des biens ont une surface totale nulle, ce qui n'est pas pertinent dans notre analyse. Nous supprimons donc ces lignes. Nous faisons cependant le choix de ne pas retraiter les valeurs extrêmes de cette colonne. En effet, nous avons déjà retraité les valeurs extrêmes des colonnes *surfaces réelles* bâtie et *surface terrain*. Nous ne souhaitons plus restreindre notre échantillon sur ce critère.

Nous visualisons la relation entre valeur foncière du bien et surface totale :



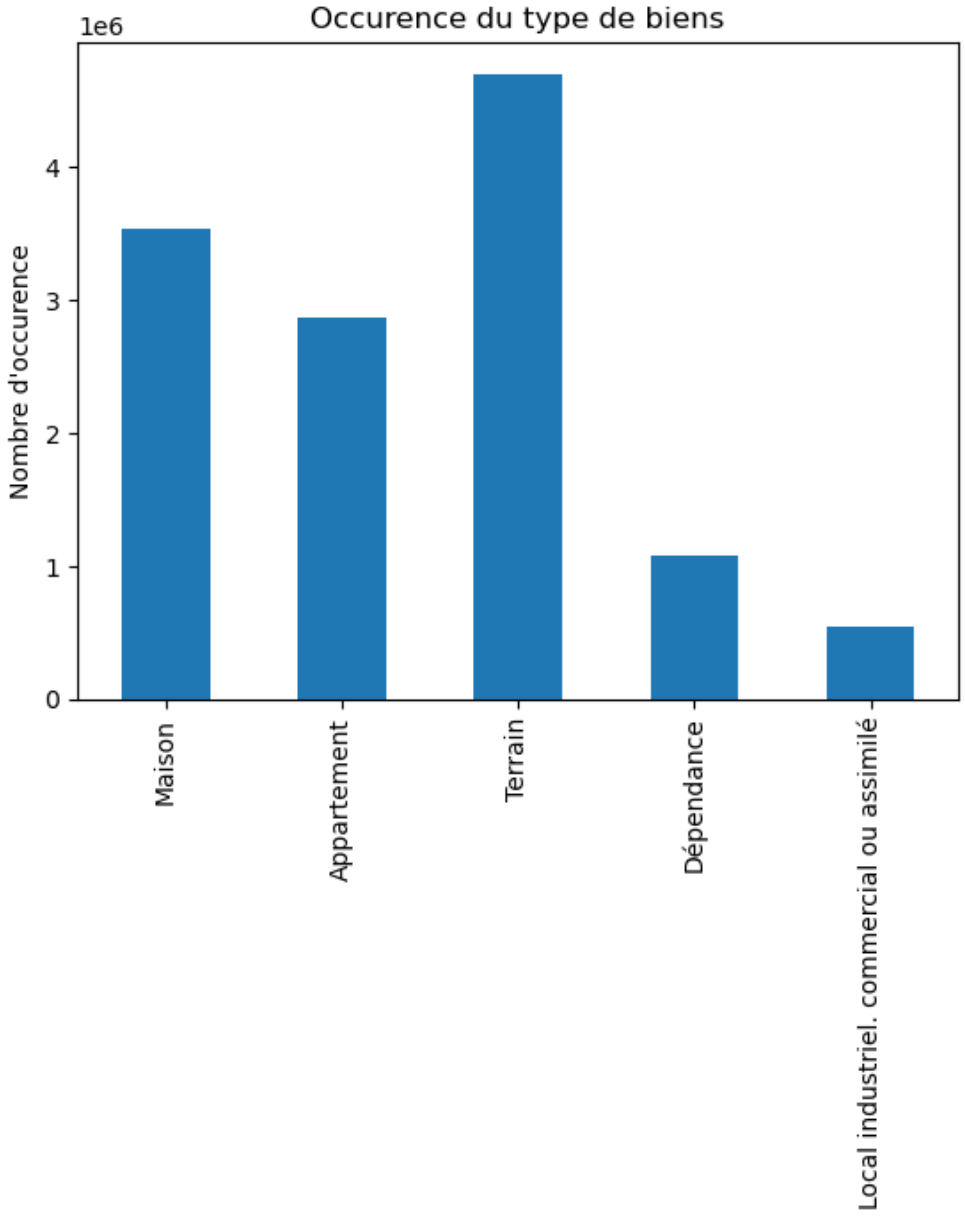
La représentation graphique nous permet de voir que l'on pourrait modéliser la relation entre valeur foncière d'un bien et sa surface totale par une fonction inverse.

3.3. Caractéristiques du bien

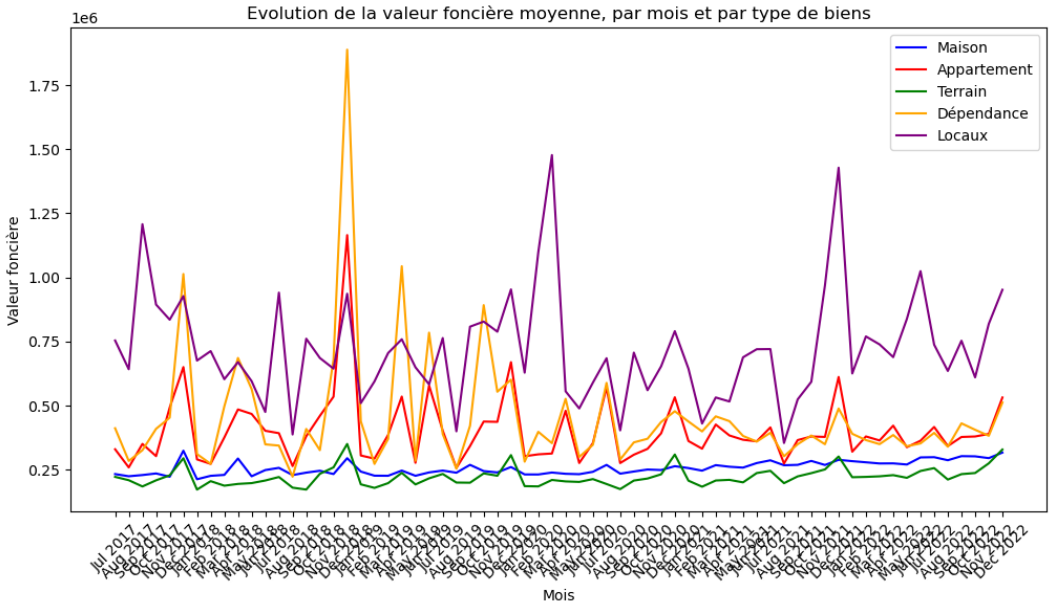
La surface d'un bien n'est pas le seul déterminant de sa valeur foncière. Nous nous intéressons maintenant à certaines caractéristiques du bien.

3.3.1. TYPE DE LOCAL

Nous visualisons le nombre de transactions par type de local :



Nous pouvons voir que la majorité des biens vendus sont des biens d'habitation (maisons et appartements confondus) puis des terrains. Les biens commerciaux étant le type de bien le moins échangé sur notre échantillon temporel.



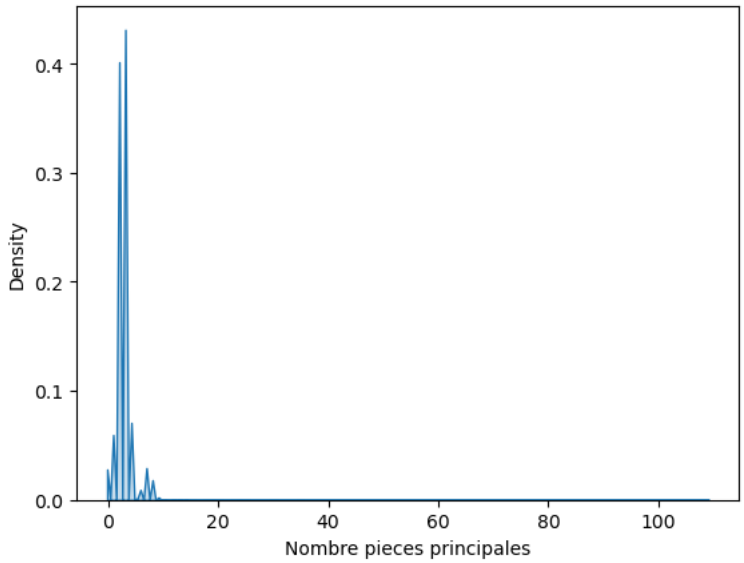
Nous pouvons voir que la valeur foncière moyenne des biens échangés dépend fortement du type desdits bien. Par exemple, la valeur moyenne d'un appartement est bien plus élevée que celle d'une maison.

De plus, la valeur moyenne des biens fluctue de façon importante dans le cas des dépendances, des appartements, et des locaux industriels et commerciaux. En revanche, nous observons peu de fluctuations dans le cas des maisons, même si les prix semblent être en constante augmentation.
Nous notons également la présence de saisonnalité dans la vente de terrains. En effet, la valeur foncière moyenne atteint son maximum à chaque fin d'année.

Nous pouvons donc dire que le type de bien influence fortement la valeur foncière des biens.

3.3.2. NOMBRE DE PIECES PRINCIPALES

<u>Fonction de densité :</u>

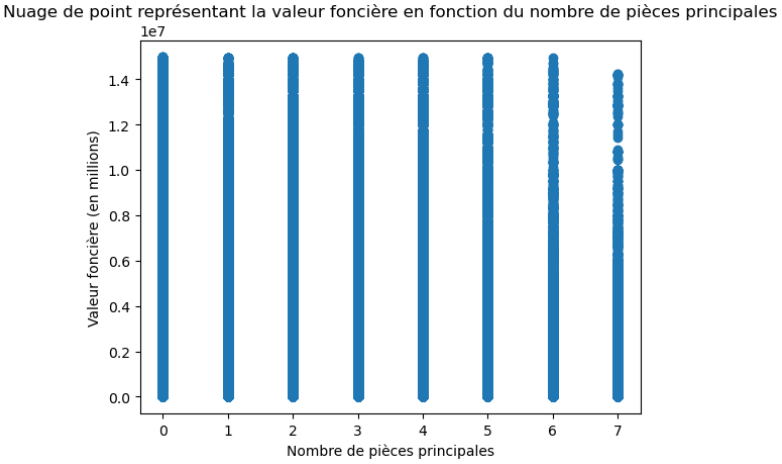


Statistiques descriptives :

	Valeur
Moyenne	1,7 pièces
Minimum	0 pièce
1er quantile	0 pièce
Médiane	1 pièce
3eme quantile	3 pièces
Maximum	109 pièces

Nous pouvons voir que la majorité des biens possède moins de 10 pièces principales si l'on exclue les quelques valeurs extrêmes (que nous retraitons). Les statistiques sont conformes à ce que l'on peut attendre de la variable. 25% des biens n'ont pas de pièces principales (terrains, studios) ; et 50% des biens ont entre 1 et 3 pièces principales, ce qui correspond à la taille standard d'un appartement ou d'une maison.
Il serait également intéressant d'avoir accès au nombre de pièces totales de la maison, c'est-à-dire une statistique qui inclurait cuisines et salles d'eau (les salles de bains, contenant une baignoire, étant en effet moins communes de nos jours).

Nous visualisons la relation entre valeur foncière et nombre de pièces principales :



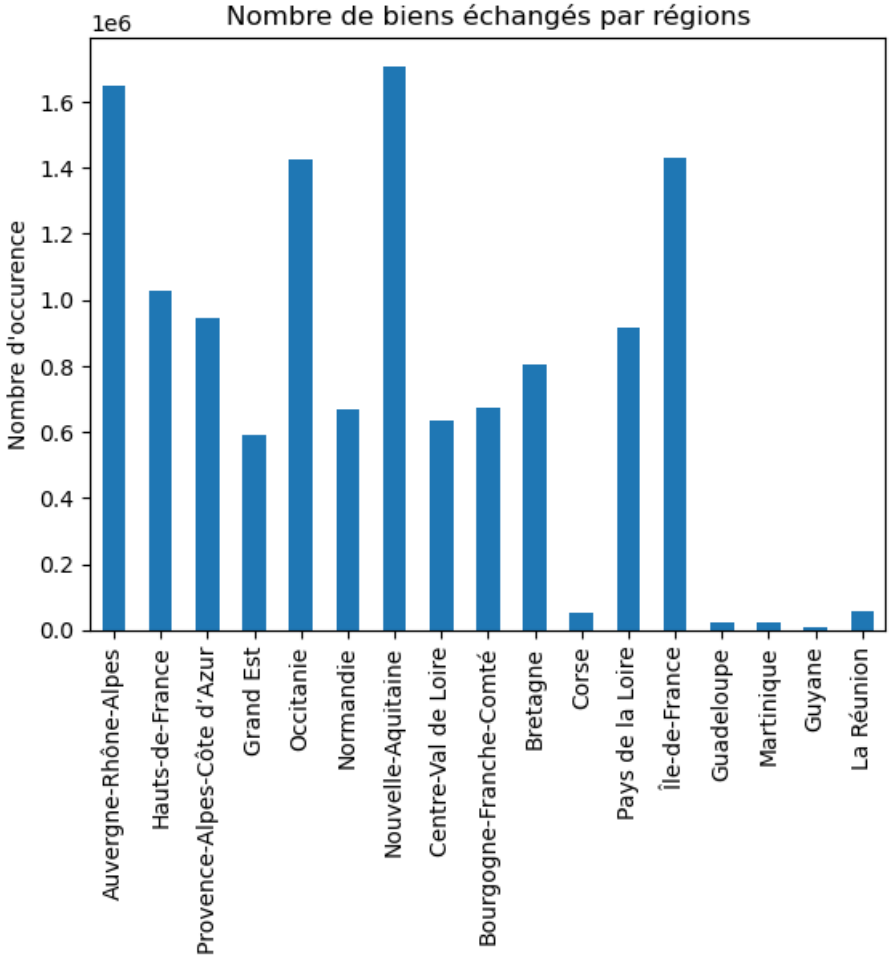
Nous pouvons voir que le nombre de pièces principales ne semble pas être un facteur d'influence de la valeur du bien. En effet, quelque soit le nombre de pièces principales, toute la gamme de valeurs foncières est disponible.
Nous pensons que, plus que le nombre de pièces principales, c'est leur surface qui est importante. Il est donc préférable d'analyser la surface d'un bien plutôt que le nombre de pièces principales.

3.4. Localisation du bien

Il est souvent considéré que la localisation d'un bien est sa caractéristique la plus importante. Nous cherchons à savoir si c'est le cas.

3.4.1. REGION OU SE SITUE LE BIEN

Nous représentons le nombre de transactions par région :



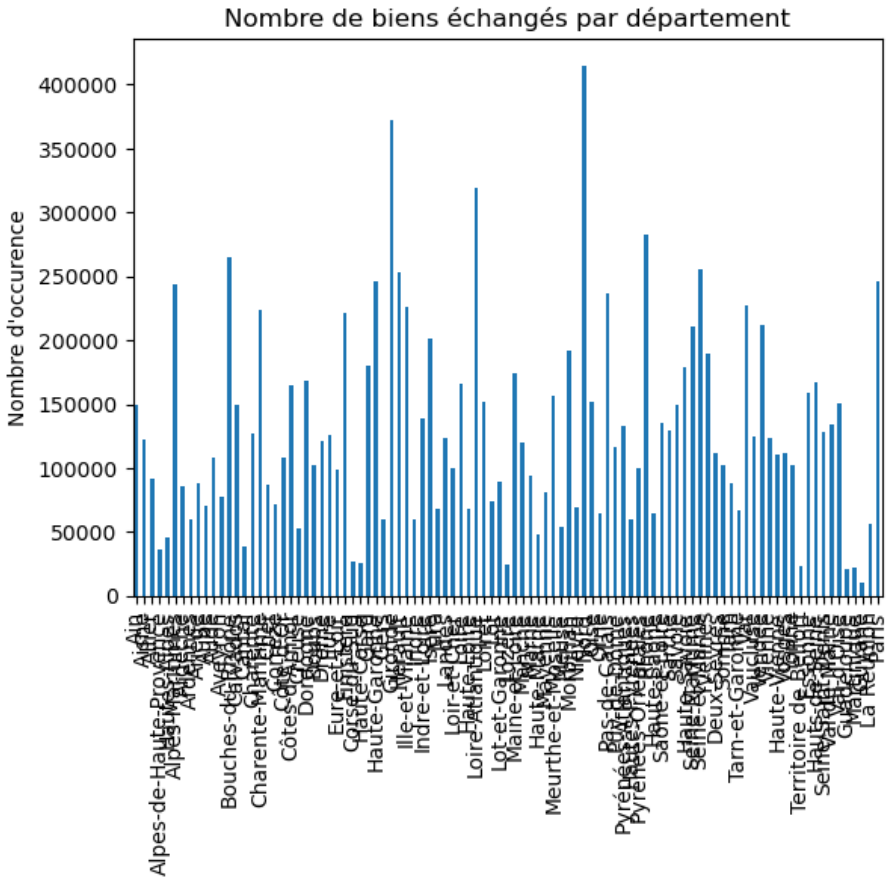
Nous pouvons voir que le nombre de transactions immobilières dépend fortement de la région dans laquelle le bien se situe. Nous émettons onc l'hypothèse que la valeur foncière d'un bien dépend de la région où il se situe.
Nous avons également effectué les graphiques de la valeur foncière moyenne mensuelle par région, qui nous permettent de confirmer cette hypothèse.

La valeur foncière d'un bien dépend donc de la région dans lequel il se situe.

Contrairement ce que l'on pouvait attendre ce n'est pas en Ile de France que l'on trouve les plus gros volumes de vente mais en Nouvelle Aquitaine.

3.4.2. DEPARTEMENT OU SE SITUE LE BIEN

La région a donc un impacte sur la valeur foncière d'un bien, mais quand est-il du département où il se situe ?



Bien que l'axe des abscisses ne soit pas lisible, nous pouvons voir que, comme pour les régions, le nombre de transaction dépend du département dans laquelle se situe le bien.
Nous avons également réalisé les 96 graphiques (un par département) représentant l'évolution de la valeur foncière moyenne mensuelle des biens.

La conclusion est identique au cas des régions : la valeur foncière d'un bien fluctue en fonction du département dans lequel il se trouve.

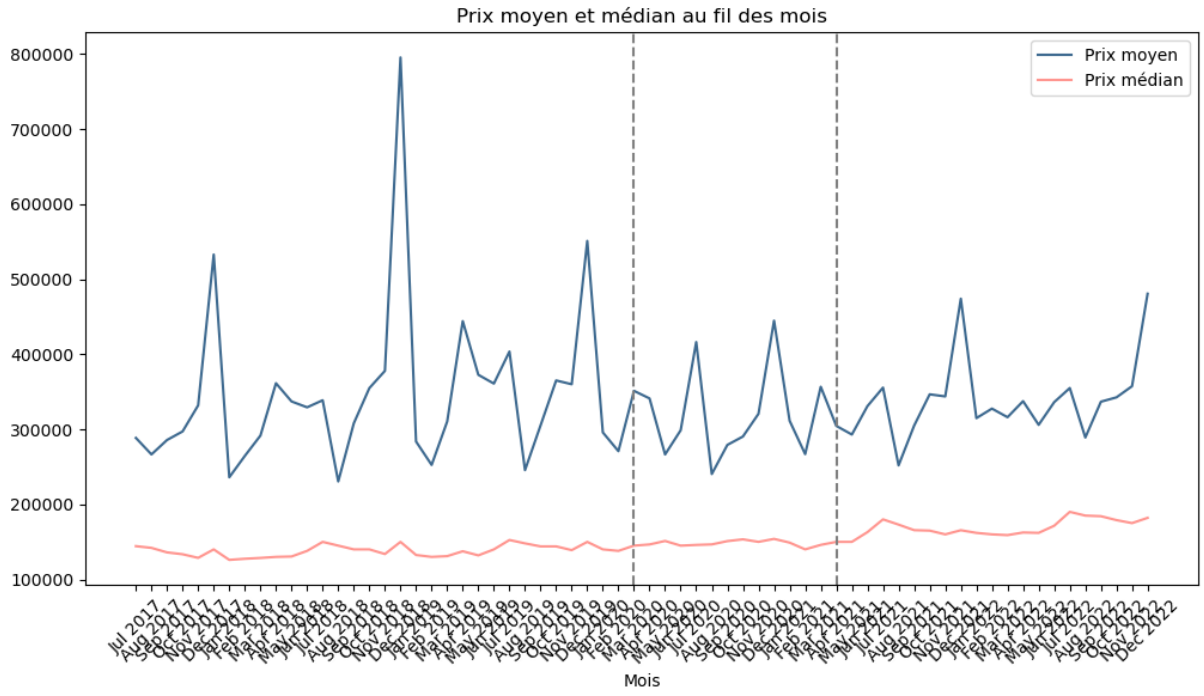
3.5. Conclusion

Nos premières analyses graphiques nous ont permis de déterminer que la surface totale, le type et la localisation du bien influencent le prix de l'immobilier.
Si nous devions établir un modèle prédictif des prix de l'immobilier, ce seraient les premières variables que nous prendrions en compte afin de prédire les prix.

4. IMPACT DE LA CRISE SANITAIRE SUR LE MARCHE DE L'IMMOBILIER FRANCAIS

Avant de débuter notre analyse, nous définissions la période temporelle sur laquelle a eu lieu la crise sanitaire. Nous considérons qu'elle a débuté le 17 mars 2020 (début du premier confinement en France) et s'est terminée le 3 mai 2021 (fin du troisième confinement).

4.1. Etude générale des prix de l'immobilier



Sur ce premier graphique, nous pouvons voir l'évolution des prix moyens et médians en France de juillet 2017 à décembre 2022. De prime abord, la crise ne semble pas avoir eut un impact significatif sur le marché de l'immobilier français. En effet, même avant la crise, le prix moyen de l'immobilier est très volatile ; on n'exclut pas la présence d'un facteur de saisonnalité dans les ventes françaises.
Bien que la crise sanitaire n'ait pas entrainé de chute sur le marché immobilier français, nous notons que le prix médian des biens vendus en France est en augmentation depuis la fin de la crise.
A première vue, l'immobilier tient bien son rang de valeur refuge en période de crise ou d'instabilité.

Nous avons effectué les mêmes graphiques pour le cas des régions, et nous pouvons voir que l'impacte de la crise sanitaire n'est pas homogène sur l'ensemble du territoire. Par exemple, nous avons les impacts suivants :

Bretagne : Bien que le volume des ventes se soit effondré, le prix moyen de l'immobilier a explosé.

Centre-Val de Loire : le prix moyen de l'immobilier est en légèrement augmentation, mais le volume des transactions est plus élevé qu'avant le Covid.

Guadeloupe : depuis la crise sanitaire, le nombre de transactions a chuté, atteignant moins de 100 transactions par mois sur plusieurs mois.

Nouvelle Aquitain : à la suite de la crise sanitaire, nous avons une hausse du nombre de transactions avec un prix moyen en hausse (qui atteint des sommets fin 2022).

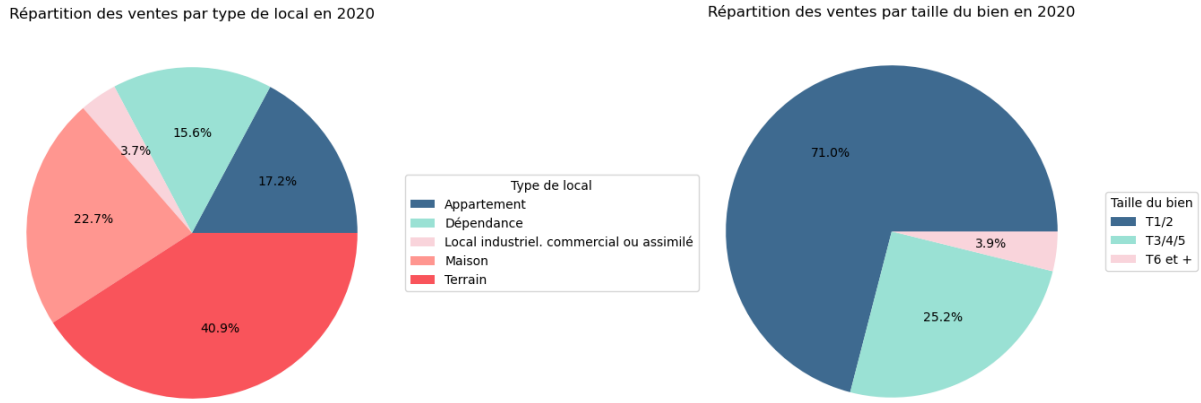
Provence-Alpes-Côte d'Azur : le nombre de ventes ne semble pas avoir été impacté par la crise, cependant nous constatons une hausse du prix moyen de l'immobilier dans la région.

Lorsque nous regardons ces mêmes statistiques pour les départements, nous arrivons à la même conclusion : l'impact de la crise sanitaire n'est pas homogène sur le territoire français, et ce même au sein d'une unique région.

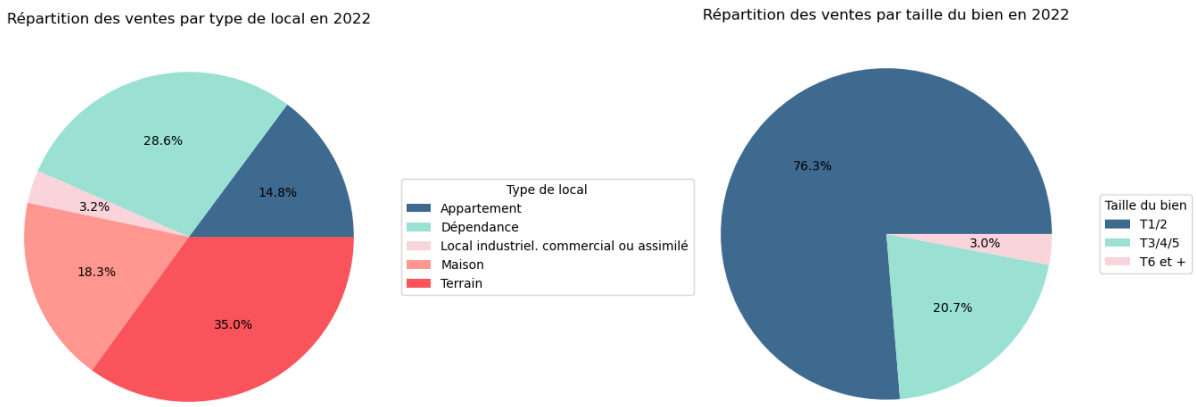
4.2. Analyse du type de biens vendus

Afin d'affiner notre analyse, nous avons voulu voir si le marché de l'immobilier avait changé suite à la crise sanitaire. Pour cela, nous avons analysée la composition des biens vendus chaque année.

Au niveau métropolitain, nous obtenons, en 2020, les statistiques suivantes :



Et en 2022 :



Ainsi, nous pouvons voir qu'entre 2019 et 2022, la part des ventes de dépendance a fortement augmenté (+ 15%). En contrepartie, ce sont les ventes de terrain (- 6,8 %), d'appartements (- 3,6%), et de maisons (- 3,9%) qui ont diminué.
Lorsque nous regardons les diagrammes circulaires, nous pouvons voir que ces changements de proportion s'opèrent entre 2020 et 2021. Ceci nous permet de déterminer que la crise sanitaire a eu un impact sur le type de biens vendus sur le marché de l'immobilier français.

Ces graphiques nous permettent également de constater que la majorité des biens vendus, quel que soit l'année sont des biens dits T1/T2, soit des biens ayant une ou deux pièces principales. Nous notons cependant qu'après la crise sanitaire, nous avons une légère augmentation (de 70 à 75%) de la proportion de biens T1 et T2 vendus.
Il semblerait donc que la crise sanitaire ait donné lieu à de plus nombreuses transactions de biens avec peu de pièces principales.

Nous regardons maintenant si ces variations se trouvent dans les mêmes proportions au niveau régional et départemental. Pour cela, nous avons effectué une fonction prenant l'année et la région étudiée et effectue les mêmes graphiques qu'au niveau national (ainsi qu'un tableau de statistiques contenant le nombre de transactions, la moyenne et la médiane des prix).

Nous commentons les mêmes régions que dans la partie 4.1. :

Bretagne : pour rappel, les prix bretons sont en constantes augmentations avec une tendance haussière pour les volumes des ventes. Le profil des biens vendus en Bretagne suit celui des biens vendus en France : 75% de T1/T2 en 2022 et le type des biens vendus corresponds également à celui de la France métropolitaine.

Centre-Val de Loire : les changements dans la structure des biens vendus suivent ceux du niveau national : hausse importante des ventes de dépendances et mêmes proportions des biens vendus en fonction de leur taille.

Guadeloupe : le nombre de transactions en Guadeloupe a diminué après la crise sanitaire, et le profil des ventes également : la majorité des ventes en 2022 étaient des biens d'habitation (plus de 50% des transactions, contre une moyenne nationale de 33% au niveau national. La

part des biens vendus reste en revanche stable dans le temps avec des ventes d'environ 25% de terrains, 25% de maisons, et 25% d'appartements.

Nouvelle Aquitaine : la grande majorité des transactions sont des terrains (55,2% en 2020, 49% en 2022) contre une moyenne nationale de 35%.

Provence-Alpes-Côte d'Azur : les changements dans la structure des biens vendus suivent ceux du niveau national : hausse importante des ventes de dépendances (mais dans des proportions plus élevées qu'au niveau national) et mêmes proportions des biens vendus en fonction de leur taille.

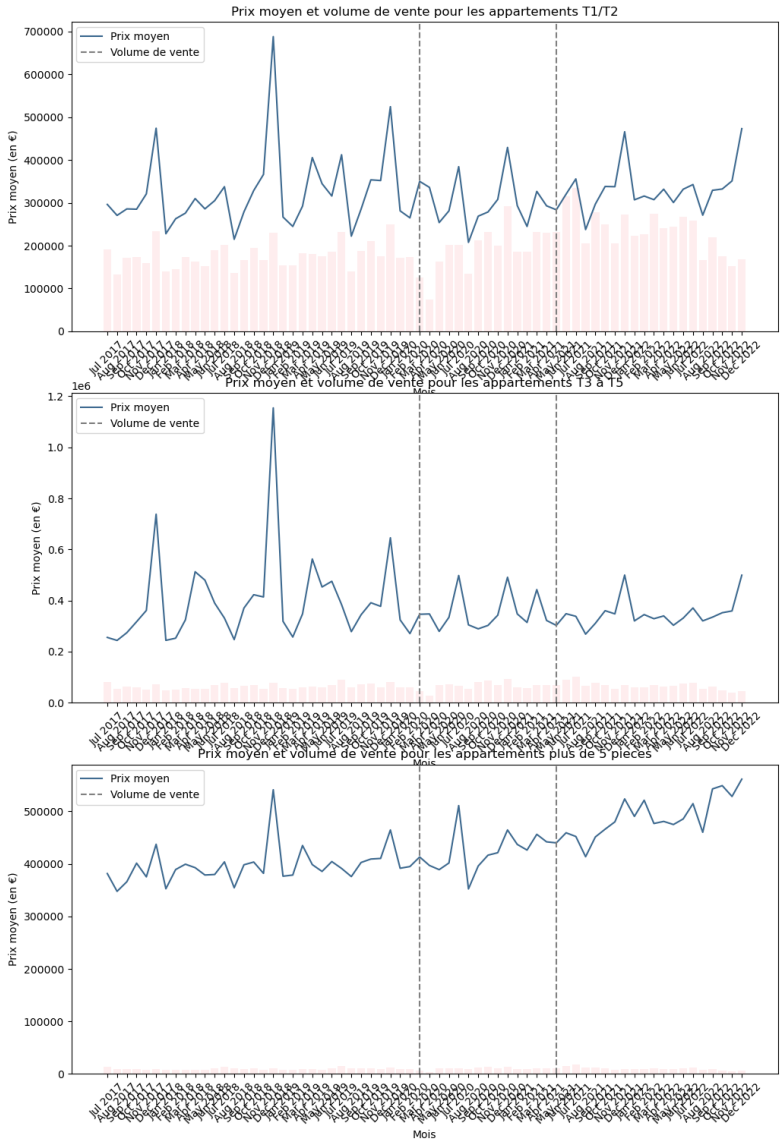
Malgré quelques changements, nous remarquons que les types de biens vendus n'ont pas particulièrement changé avec la crise sanitaire. Cependant, dans leur majorité la valeur foncière des biens est en hausse.

Cette hausse des prix est imputable à une hausse très importante de la demande par suite de la crise sanitaire. En effet, une partie des acheteurs a délayée l'achat ou la vente de leur bien à cause de la crise. A la fin de la crise, le marché de l'immobilier a donc continué à plein régime. Cette hausse est exacerbée en 2022 avec la hausse des taux d'intérêts et les difficultés économiques liées à la guerre en Ukraine.

Nous avons également effectué une fonction nous permettant d'analyser ces variables au niveau départemental. Nos conclusions sont identiques.

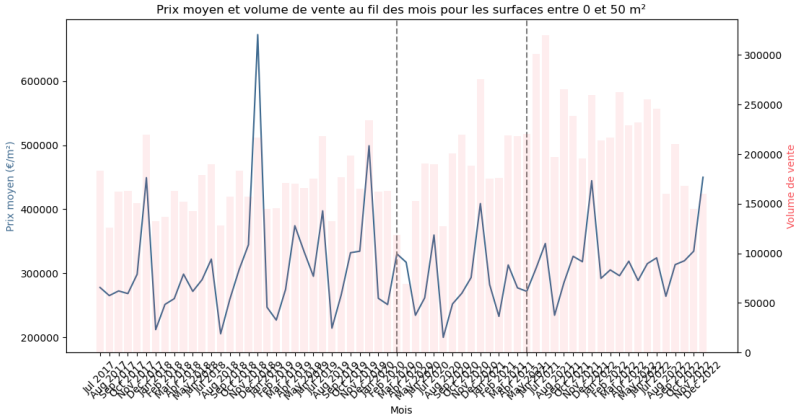
4.3. Analyse des prix de l'immobilier en fonction de la surface ou du nombre de pièces principales

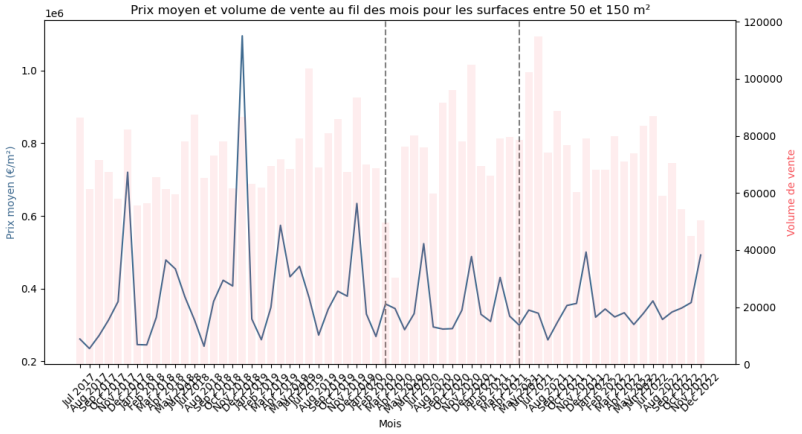
Nous analysons les prix des biens en fonction du nombre de pièces principales :

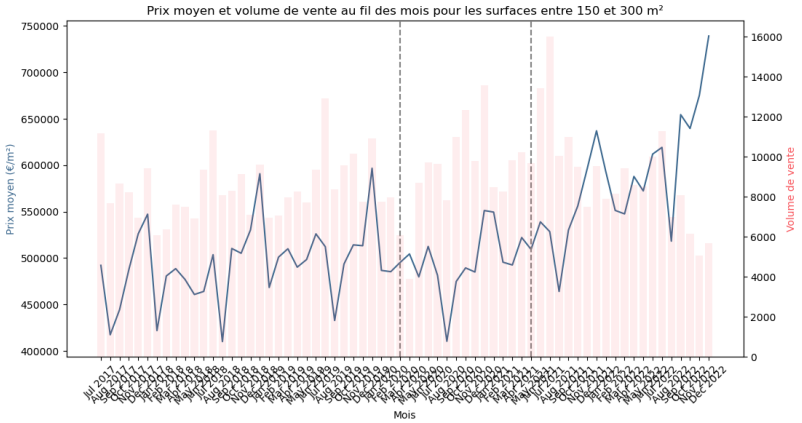


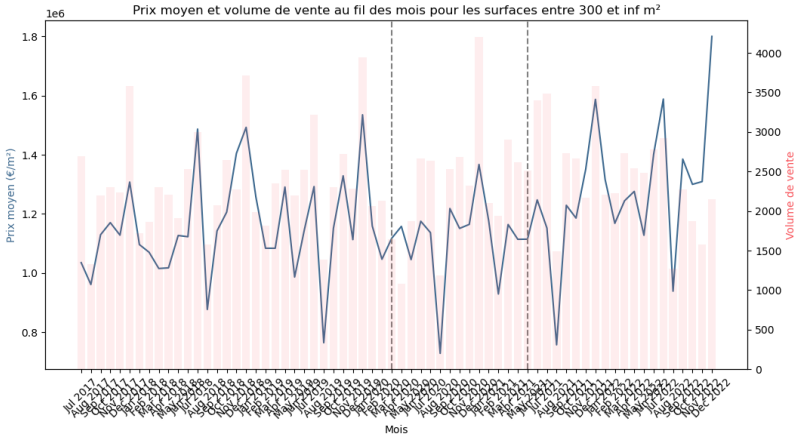
Nous pouvons voir que dans leur grande majorité les transactions se réalise pour des biens ayant une à deux pièces principales (cf. volume des ventes en rose).
Graphiquement, la crise sanitaire ne semble pas avoir eu un impact particulier sur le prix moyen des biens vendus. Nous remarquons juste une hausse des prix depuis mai 2021 pour les appartements de plus de 5 pièces principales. Cependant, nous ne pouvons rien en conclure à cause du faible nombre de transactions.

Nous analysons maintenant le prix de l'immobilier en fonction de la surface du bien :









Dans chaque cas, les prix des biens est très volatile d'un mois sur l'autre. Nous pensons donc qu'il existe un facteur de saisonnalité dans la détermination des prix de biens.

Cette forte volatilité nous empêche d'être conclusif sur l'impact de la crise sanitaire sur la valeur foncière des biens, excepté dans le cas des biens ayant une surface entre 150 et 300 m², dont la valeur est en très forte augmentation depuis la fin de la crise sanitaire.

4.4. Conclusion

En conclusion, la crise sanitaire n'a pas eu d'impact particulier sur les prix de l'immobilier français en général. Nous avons néanmoins pu constater que l'impact de la crise sanitaire n'est pas homogène sur l'ensemble du territoire avec certains territoires qui sont bien impactés que d'autre, que ce soit en termes de transaction ou de valeur foncière des biens.

Nous pensons donc que d'autres variables, qui ne sont pas présentes dans notre étude influencent les prix de l'immobilier tels que les taux d'intérêts, les anticipations d'inflation des agents, et les effets de popularité d'une localisation ou d'un type de bien particulier (par exemple, une piscine augmente le prix d'un bien, mais ce n'est pas mesuré ici).

5. MODELE PREDICTIF DES PRIX DE L'IMMOBILIER FRANÇAIS

Il existe plusieurs manières de créer un modèle prédictif. Ici, nous allons droit au but en utilisant un modèle linéaire en coupe instantanée.
L'idéale aurait été d'effectuer un modèle de séries temporelles, en stationnarisant des séries afin de déterminer le modèle ARIMA ou VAR approprié. Cela voudrait dire que l'information du prix se trouve en "elle-même" et que l'on pourrait faire abstraction des variables exogènes, ce qui ne peut pas être le cas ici.

Nous avons essayer d'entrainer notre modèle sur 90% de la base et la tester sur les 10% restants mais les résultats n'étaient pas probant donc nous avons utiliser l'ensemble de la base pour avoir un modèle pertinent.

Autre information, pour notre régression linéaire sur prix de nos variables nous allons avoir besoins de variable indicatrice pour nos colonnes non numérique comme le type de local (ou la région). Ceci va se traduire par la création d'autant de colonne qu'il n'y a de type de local (ou de régions) et donc les valeurs prises par nos lignes seront 1 ou 0.

5.1. Régressions simples

OLS Regression Results

Dep. Variable:	Valeur fonciere	R-squared:	0.004
Model:	OLS	Adj. R-squared:	0.004
Method:	Least Squares	F-statistic:	2.634e+04
Date:	Tue, 18 Apr 2023	Prob (F-statistic):	0.00
Time:	20:14:36	Log-Likelihood:	-2.7335e+08
No. Observations:	17946450	AIC:	5.467e+08
Df Residuals:	17946446	BIC:	5.467e+08
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.127e+05	287.356	1088.271	0.000	3.12e+05	3.13e+05
Surface reelle bati	50.9386	0.378	134.768	0.000	50.198	51.679
Nombre pieces principales	1.604e+04	120.030	133.609	0.000	1.58e+04	1.63e+04
Surface terrain	3.8976	0.019	208.023	0.000	3.861	3.934

Omnibus:	25026762.265	Durbin-Watson:	0.118
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5683809141.932
Skew:	8.404	Prob(JB):	0.00
Kurtosis:	88.548	Cond. No.	1.60e+04

Notre premier modèle est simple : expliquer les prix de l'immobilier en fonction de la surface bâtie, surface du terrain, et nombre de pièces principales.
Bien que nos trois variables soient significativement différentes de 0, nous pouvons voir que notre modèle est peu explicatif. En effet, notre R^2 est de 0,4%.
Ce n'est pas un résultat surprenant. En effet, nous avons établis dans les parties précédentes que la valeur d'un bien fluctuait beaucoup en fonction de l'endroit où il se trouvait. De plus, nous avons pu voir que les prix d'un bien ne sont pas constants d'un an sur l'autre, avec par exemple de fortes augmentations de prix entre 2020 et 2022. Enfin, nous avions également vu que le type de bien (terrain, appartement ou maison par exemple) influençait la valeur d'un bien.
Or, ce sont trois variables que nous n'avons pas pris en compte dans notre régression. Il était donc attendu que celle-ci soit mauvaise.

OLS Regression Results

Dep. Variable:	Valeur fonciere	R-squared:	0.029
Model:	OLS	Adj. R-squared:	0.029
Method:	Least Squares	F-statistic:	7.531e+04
Date:	Tue, 18 Apr 2023	Prob (F-statistic):	0.00
Time:	20:15:39	Log-Likelihood:	-2.7313e+08
No. Observations:	17946450	AIC:	5.463e+08
Df Residuals:	17946442	BIC:	5.463e+08
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Surface reelle bati	30.8633	0.376	82.037	0.000	30.126	31.601
Nombre pieces principales	5.509e+04	271.456	202.952	0.000	5.46e+04	5.56e+04
Surface terrain	4.8148	0.019	258.391	0.000	4.778	4.851
Type local_Appartement	3.859e+05	901.764	427.964	0.000	3.84e+05	3.88e+05
Type local_Dépendance	3.935e+05	528.868	743.972	0.000	3.92e+05	3.94e+05
Type local_Local industriel. commercial ou assimilé	8.437e+05	1233.535	683.978	0.000	8.41e+05	8.46e+05
Type local_Maison	3.433e+04	1256.726	27.318	0.000	3.19e+04	3.68e+04
Type local_Terrain	2.099e+05	377.424	556.099	0.000	2.09e+05	2.11e+05

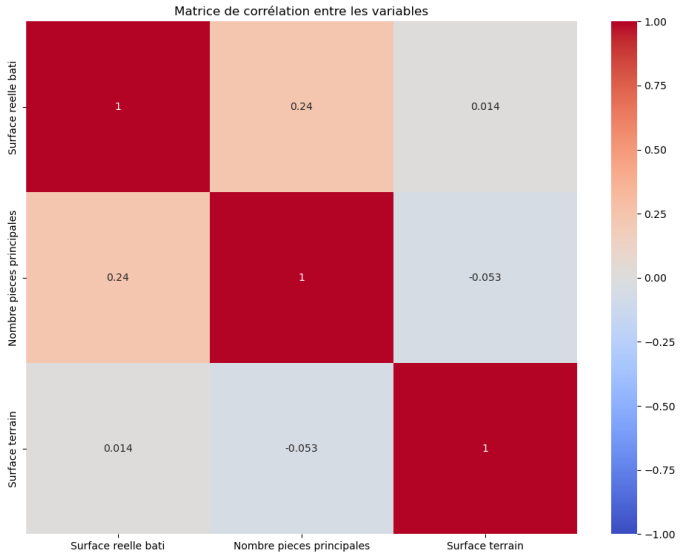
Omnibus:	24870879.899	Durbin-Watson:	0.146
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5547338137.986
Skew:	8.308	Prob(JB):	0.00
Kurtosis:	87.513	Cond. No.	8.05e+04

Tout d'abord, nous retirons la constante de notre modèle. En effet, nous utilisons des variables indicatrices pour désigner le type de bien vendus et ajouter une constante créerait une combinaison linéaire entre celle-ci et les variables indicatrices ; ce qui fausserait notre modèle. Nous ajoutons donc le type de local à notre modèle précédent.
Nous pouvons tout de suite voir que toutes nos variables sont significatives et que notre modèle est plus explicatif auparavant avec un R^2 de 2,9%.

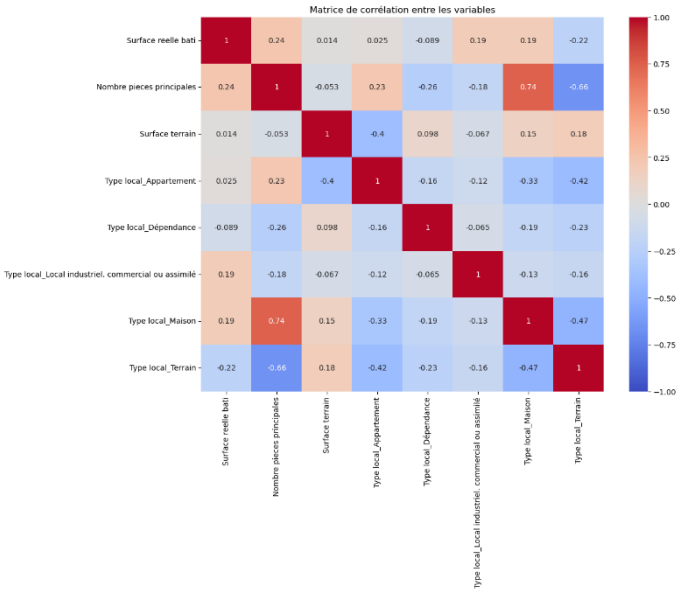
Nous avons ensuite souhaité ajouter les régions à notre modèle. Cependant, étant donné que nous avons 13 régions et 5 types de locaux, le nombre de données était trop importantes, et il était impossible d'effectuer la régression.
C'est décevant car nous pensons qu'ajouter les régions aurait significativement amélioré le modèle.

5.2. Corrélation des variables

Se pose maintenant la question d'autocorrélation de notre modèle. Il est possible que certaines variables soient corrélées entre elles ce qui fausserait notre régression.

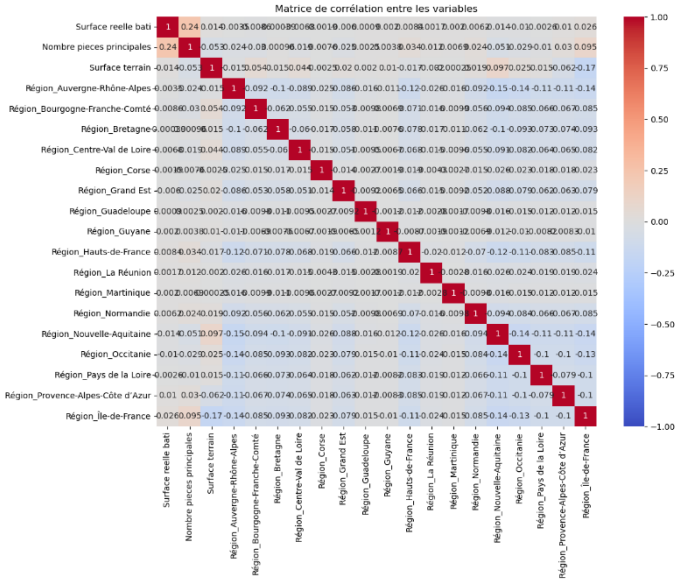


Nous pouvons déjà voir que les variables de notre premier modèle ne sont pas corrélées.



Nous ajoutons maintenant les types de locaux à notre matrice de corrélation. Nous pouvons voir que les variables de notre second modèle sont corrélées entre elles.
Ainsi, le nombre de pièces principales est corrélé a 0,74 avec les maisons ; et à -0,66 avec les terrains.
Nous avons donc un problème de corrélation qui fausse notre modèle économétrique.

Nous regardons maintenant si nous obtenons de la corrélation avec les régions uniquement.



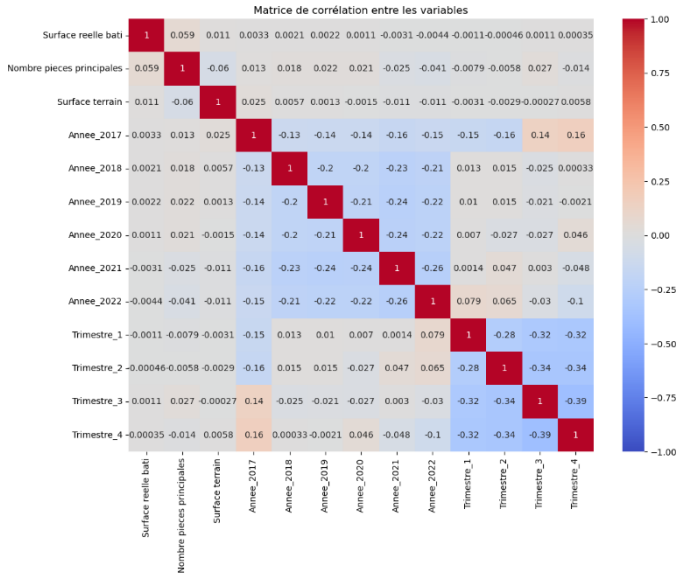
Nous pouvons voir que nous n'avons aucun problème de corrélation dans le cas des régions. Cependant, effectuer une régression avec les régions en temps que variable explicative requiert trop de mémoire de la part de nos ordinateurs ; cela ne fonctionne pas, ce qui est dommage.

5.3. Prise en compte de l'année de vente

Etant donné que nos deux précédents modèles sont inadaptés car trop lourd ou à cause de corrélation entre les variables, nous ajoutons la troisième variable explicative mentionnée

dans la partie 5.1. : la date de vente du bien. Nous ajoutons donc l'année et le trimestre de vente du bien afin de voir si cela améliore notre modèle.

Nous commençons par regarder la corrélation entre nos variables explicatives.



Nous pouvons voir que les trimestres sont corrélés entre eux. Nous faisons donc le choix de les retirer de notre modèle.

OLS Regression Results

Dep. Variable:	Valeur fonciere	R-squared:	0.005
Model:	OLS	Adj. R-squared:	0.005
Method:	Least Squares	F-statistic:	1.044e+04
Date:	Tue, 18 Apr 2023	Prob (F-statistic):	0.00
Time:	21:30:15	Log-Likelihood:	-2.7334e+08
No. Observations:	17946450	AIC:	5.467e+08
Df Residuals:	17946441	BIC:	5.467e+08
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.683e+05	255.591	1049.619	0.000	2.68e+05	2.69e+05
Surface reelle bati	50.9096	0.378	134.707	0.000	50.169	51.650
Nombre pieces principales	1.592e+04	120.207	132.436	0.000	1.57e+04	1.62e+04
Surface terrain	3.8903	0.019	207.563	0.000	3.854	3.927
Annee_2017	3.869e+04	721.131	53.645	0.000	3.73e+04	4.01e+04
Annee_2018	6.619e+04	539.496	122.681	0.000	6.51e+04	6.72e+04
Annee_2019	6.328e+04	520.899	121.487	0.000	6.23e+04	6.43e+04
Annee_2020	2.21e+04	525.935	42.012	0.000	2.11e+04	2.31e+04
Annee_2021	3.519e+04	475.744	73.974	0.000	3.43e+04	3.61e+04
Annee_2022	4.283e+04	501.587	85.393	0.000	4.18e+04	4.38e+04

Omnibus:	25015437.479	Durbin-Watson:	0.118
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5670655918.934
Skew:	8.397	Prob(JB):	0.00
Kurtosis:	88.448	Cond. No.	4.71e+15

Notre modèle est toujours peu explicatif (R^2 = 0,5%) malgré que toutes nos variables soient significatives.
Cependant les coefficients ont tous les signes attendus (positif). En effet, plus un bien a une grande surface ou un nombre de pièces principales élevées, plus il est cher.
Enfin, nous pouvons voir l'impact de la crise du covid sur nos variables. En effet, entre 2019 et 2020, le coefficient est divisé par 3. Cependant, il augmente à nouveau une fois que la crise sanitaire est passée.

Conclusion

Nous avons pu tester plusieurs modèles économétriques pour prédire les prix de l'immobilier français. Cependant, nous avons fait face à plusieurs problèmes : corrélation des variables, manque de puissance des ordinateurs ou modèles peu explicatifs.

Le meilleur modèle que nous avons pu établir est le troisième. La valeur foncière des biens est expliquée par : l'année de vente, la surface réelle bâtie, la surface du terrain et le nombre de pièces principales.
Bien que peu explicatif, nous n'avons pas de problème de corrélation entre les variables comme notre second modèle (type de local au lieux de l'année de vente) qui était plus explicatif mais avait ce problème.

CONCLUSION

Nous avons pu étudier l'impact de la crise sanitaire sur les prix de l'immobilier sur le marché français au niveau national, régional, et départemental.

Nous avons conclu que l'impact de la crise sanitaire n'est pas homogène sur le territoire ni sur le type de biens vendus. En effet, certaines régions ont été plus impactées que d'autres et certains types de biens ont vu leur prix augmenter plus fortement.
Cependant, il aurait été intéressant de se focaliser sur une commune pour voir précisément les impacts que la crise sanitaire a eu sur celle-ci. De plus, se focaliser sur une unique commune permettrait de réduire considérablement le nombre de données et donc d'ajouter des variables explicatives.

Nous avons enfin établi un modèle économétrique pour prédire les prix futurs de l'immobilier en utilisant un modèle linéaire. Notre modèle est correct dans sa démarcher mais peu explicatif.
Ce résultat n'pas surprenants. En effet, nous savons qu'une multitude de facteurs externes influencent les prix de l'immobilier tels que les taux d'intérêt, les conditions d'accès à des prêts immobiliers, les anticipations d'inflation, la taille d'une ville ou encore la popularité d'une localisation.