RAPPORT DE PROJET STATISTIQUE

Evolution du prix de l'immobilier en France

Chaimaa Beladraoui
Maxime Bonnet
Tasnîme-Jenna Louartani

Desigeo
Décembre 2022

I) Introduction

Dans le cadre du module rassemblant l'Analyse Spatiale, l'Apprentissage, les Probabilités et les Statistiques, nous nous sommes intéressés à l'analyse des données de la valeur foncière en France. Nous étudierons donc au cours de ce projet l'évolution du prix de l'immobilier en France. Puis nous comparerons les données de valeur foncière avec des données issus de critères sociaux et économiques. La problématique que nous suivrons se concentrera sur la compréhension de l'évolution du prix de l'immobilier en France.

II) Analyse des données

A) Présentation des données

Nous avons utilisé la base de données DVF (Demandes de Valeurs Foncières) d'Etalab fournie par la DINUM. Les métadonnées sont aussi présentes sur le site du Ministère de l'Economie, des Finances et de la Souveraineté industrielle et numérique.
Afin de permettre des analyses sur l'évolution, nous avons pris 5 tables de la DVF allant de 2017 à 2021 pour toute la France. Dans ces bases, il y a des informations sur la localisation et la valeur foncière des biens, la date et nature de la mutation (Vente, Vente terrain à bâtir, Vente en l'état futur d'achèvement…) et le type de bien (Local, Maison, Dépendance, Appartement…).

B) Problème dans la qualité des données

Cette base n'est pas parfaitement complétée pour pouvoir l'utiliser sans avoir des problèmes dans les résultats. En effet, certaines informations ne sont pas renseignées :

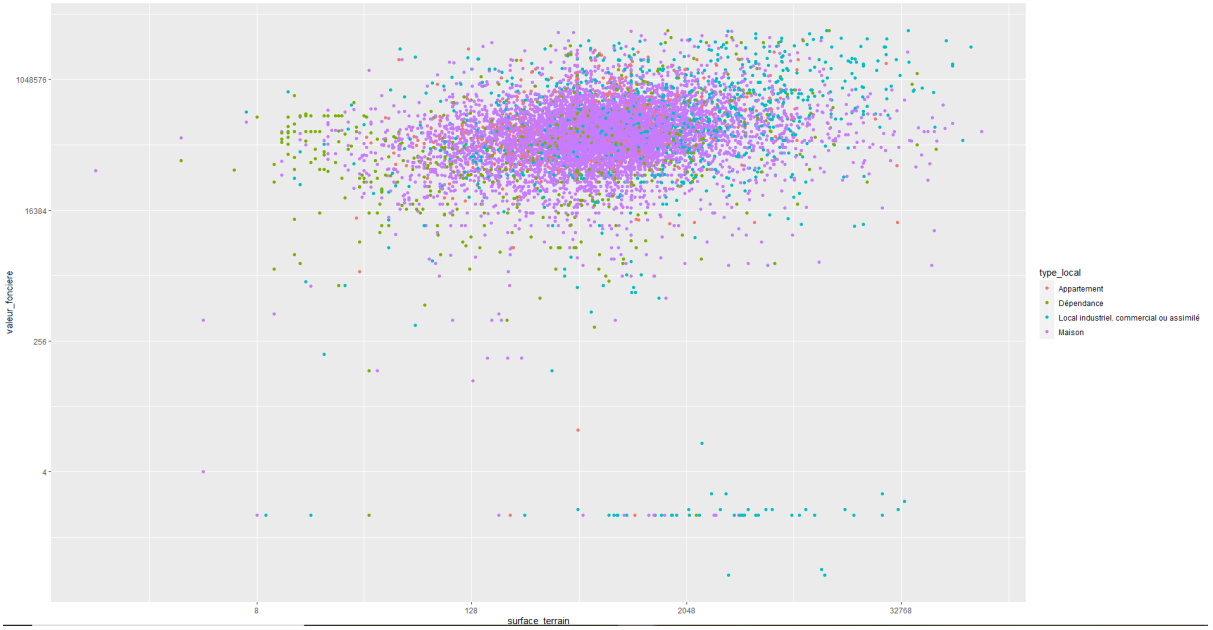
- les données de l'Alsace et la Moselle ne sont pas dans la base DVF
- certains champs ont la valeur "NA"
- valeurs totalement aberrantes : maisons à 1€, surface terrain de 1m²
- présence de doublons
- erreurs sémantiques (mauvaise localisation des parcelles…)

III) Etude des données

A) Étude préliminaire des données

1) Relation entre la valeur foncière et la surface du terrain.
Ces deux valeurs quantitatives sont les plus parlantes de la base DVF.
Une rapide visualisation nous montre de très nombreuses valeurs aberrantes (annexe). Mais lorsque l'on réaffiche le graphique avec les outliers en moins il y a toujours des valeurs qui tirent vers les extrêmes. D'où l'idée de passer en échelle logarithmique pour

ne pas supprimer encore et encore des valeurs. Après avoir défini un seuil maximal raisonnable pour le prix du bien et sa surface le résultat donne:



Valeure foncière en fonction de la surface du terrain (type de surface indéterminée non montrée dans ce graphique)

Le résultat nous montre que les valeurs sont regroupées dans une zone assez limitée, entre 16 000 et 1 000 000 d'euros et entre 32 et 16 000 mètres carrés. Ce n'est donc pas une zone si restrictive que ça.
Il aurait été intéressant de faire une estimation par noyau d'une densité ou de simplement calculer le barycentre de ces points.
On remarque sur ce graphique tout une ligne inférieure qui correspond aux biens vendus à seulement 1 euros, valeur probablement fausse indiquée par défaut dans les comptes rendus de notaire.

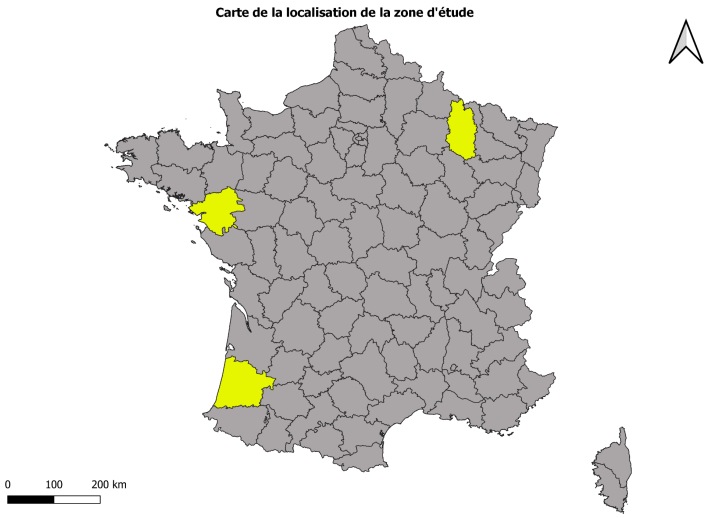
2) Travail avec toutes les années

Nous avons la chance d'avoir les valeurs des six dernières années alors nous nous sommes demandé ce qui pouvait être intéressant d'en tirer. Nous avons alors cherché quels étaient les biens, en particulier les maisons qui ont été vendues au moins 6 fois en 6 ans. Nous avons été extrêmement surpris de voir plus de 2000 transactions répondant à ces critères. En nous penchant au cas par cas sur ces maisons nous nous sommes rendu compte que la réalité était plus complexe que ça. Les parcelles correspondantes sont composées d'une multitude de maisons, soit des campings et des lotissements.
Les campings correspondent aux maisons situées sur le littoral et les lotissements ceux situés dans les grandes villes (Paris, Lyon,Toulouse, Bordeaux et Nantes).
La méthode par recherche d'unique identifiant parcellaire s'est révélée infructueuse mais nous avons tout de même pu détecter la localisation de résidences et de campings par la simple recherche de fréquence de ventes de maisons.
Nous avons également tenté la recherche avec une unique adresse sauf que les campings possèdent une même adresse pour toutes leurs maisons. Avec du temps nous aurions même pu identifier uniquement les campings par cette méthode et trouver les lotissements par recherche complémentaire de l'intersection avec la première méthode.

En comparant la carte avec celle non filtrée sur les maisons vendues on s'aperçoit qu'une énorme concentration de bien est présente dans les Alpes mais pas en tant que maison vendue. Cela s'explique par des chalets contenant de très nombreux appartements.

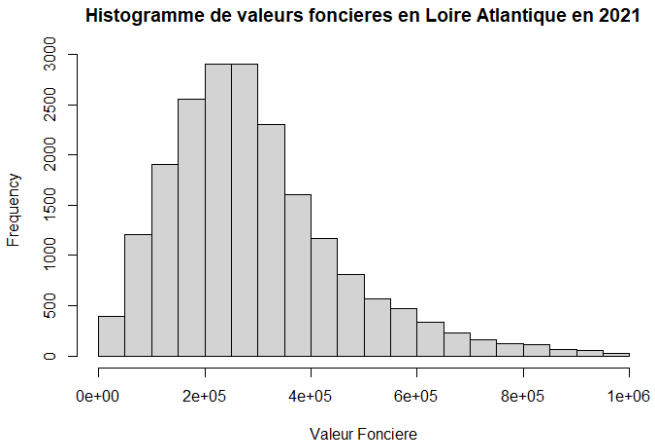
B) Présentation de la zone d'étude

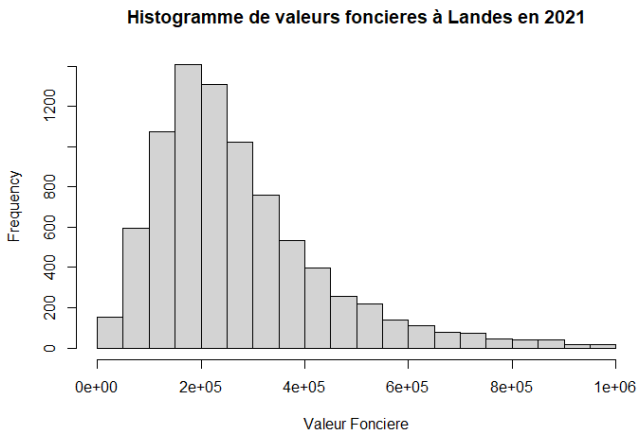
Notre zone d'étude a été réalisée dans trois départements différents : Les Landes , La Meuse et La Loire Atlantique.
Nous avons choisi le département des Landes en Nouvelle-Aquitaine en tant que département touristique connu pour ses nombreuses stations balnéaires et ses plages de sable fin ainsi que ses vastes forêts de la région.
La Loire Atlantique présente de nombreux atouts qui en font aujourd'hui un territoire attractif, il se caractérise par une forte dynamique démographique et économique tandis que la Meuse connaît un déclin démographique.



A) Evolution de la valeur foncière

Nous avons mené une étude comparative de l'évolution de la valeur foncière des ventes de maisons en 2021 entre les trois départements.

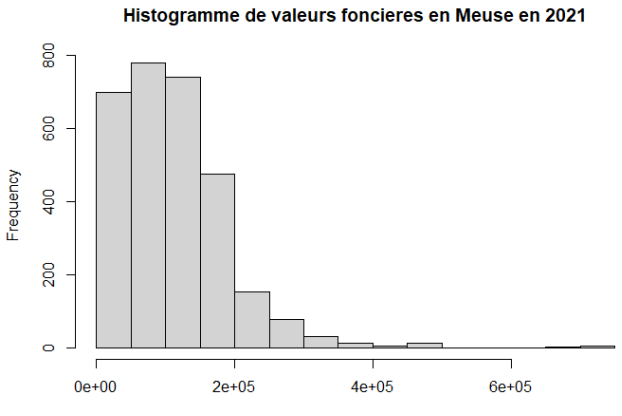




Pop : 1,426 million INSEE

Pop : 409 325 INSEE

La distribution de la valeur foncière dans les deux
départements des Landes et de La Loire Atlantique est quasiment identique .
En effet, les deux graphiques ci-dessus montrent une dispersion de la valeur foncière qui s'étend de 100 000 à 1 000 000 euros environ et une valeur foncière « typique » pour les deux distributions qui se situe entre 200 000 euros et 300 000 euros en raison du pic accentué dans cette zone.
La valeur foncière la plus fréquente est plus élevée en Loire Atlantique que dans les Landes et ce, proportionnellement au nombre d'habitants .
La similitude de la répartition de la valeur foncière dans ces deux départements peut s'expliquer par la ressemblance géographique, voire urbaine, des deux départements.



Pop : 184 474 INSEE

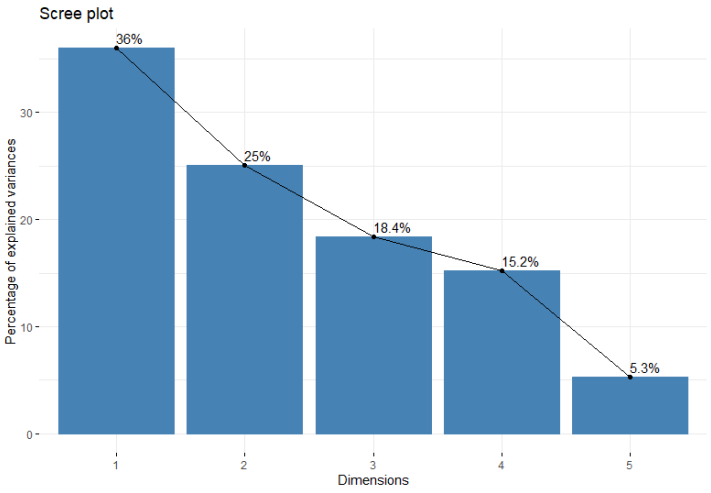
En revanche, la dispersion de la valeur foncière dans la Meuse varie d'environ 100 000 euros à 500 000 euros et une valeur foncière typique de l'ordre de 100 000 euros.

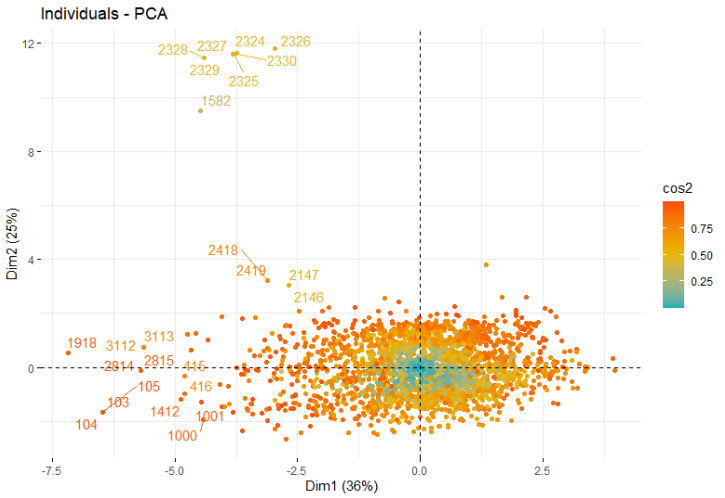
B) Analyse en composantes principales

Afin de quantifier la liaison entre les variables quantitatives des données DVF de 2017 dans les trois départements que nous avons choisis au préalable, nous avons effectué une analyse factorielle des cinq variables quantitatives suivantes : la valeur foncière, la surface réelle du bâtiment, le nombre de pièces principales ,la latitude et la longitude.

1) Département des Landes

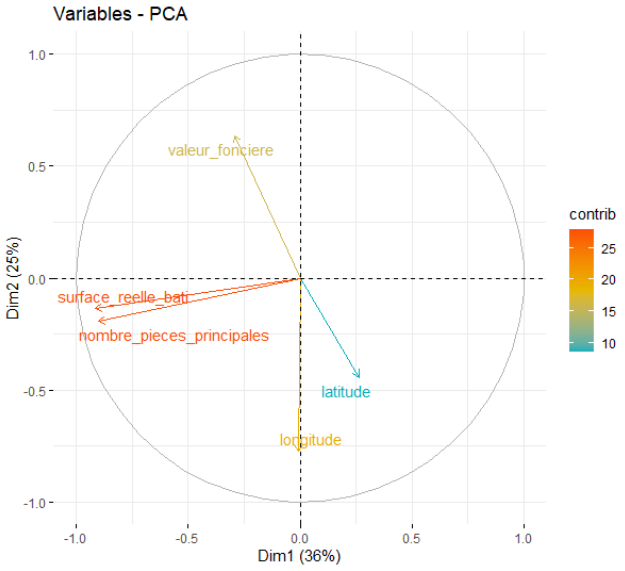
Pour commencer, nous avons appliqué l'analyse en composantes principales aux données du département des Landes en les normalisant afin que les variables soient comparables.
Le scree plot montre la proportion d'inertie capturée par les différentes composantes. Les deux premières composantes principales expliquent 61% de variation contenue dans le jeu de données . On projette les individus dans l'espace d'arrivée composé des deux premières composantes principales : plus la valeur de cos2 est élevée, mieux l'individu est représenté.





Nous avons projeté aussi les cinq variables dans l'espace d'arrivée comme le montre la figure suivante :

Les trois variables (valeur foncière, surface réelle bâti et nombre pièces principales) sont bien représentées. Il y a une forte corrélation entre les deux variables surface réelle bâti et nombre pièces principales.
En revanche, il n'y a aucune corrélation entre ces deux variables et la valeur foncière.
Il y a une décorrélation entre la valeur foncière et la latitude, ce qui met en évidence que plus on est dans le Nord ouest du département des Landes, plus la valeur foncière augmente.

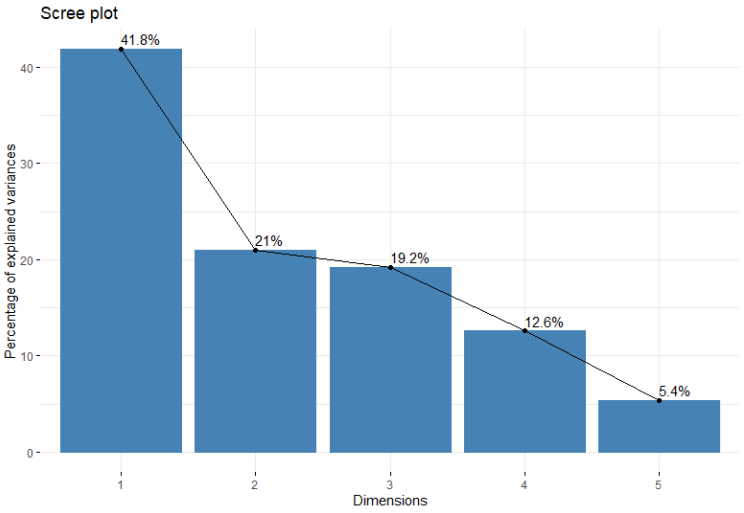


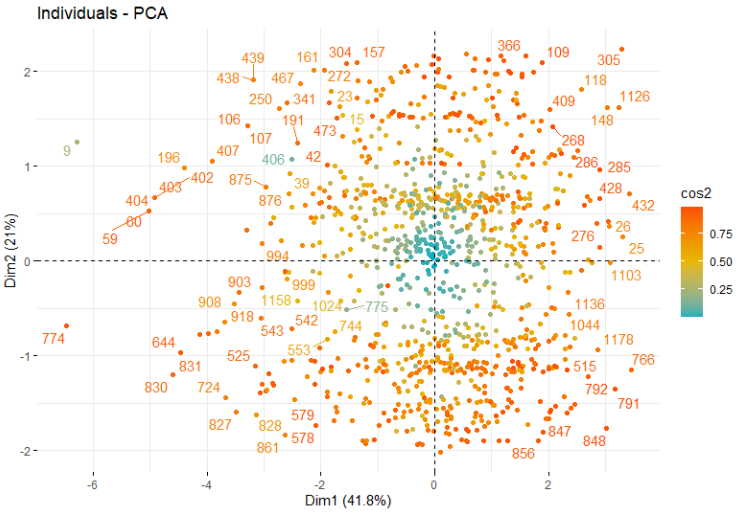
Pour mieux visualiser la dispersion et la centralité de la distribution des valeurs associés à la valeur foncière, nous avons réalisé le graphique des boxplots **(voir figure 1 en Annexe).**

D'après le graphique,il y a une grande dispersion de la valeur foncière pour les intervalles de surface de 141 à 204 m². Plus la surface augmente, plus la valeur foncière augmente sauf qu'au-delà d'une surface de plus de 236 m², la valeur foncière diminue. Cela s'explique par le fait que les maisons sont éloignées du centre ville et donc même avec une grande surface, le prix reste plus bas qu'en centre-ville.**(Voir figure 2 en Annexe)**

2) Département de la Meuse

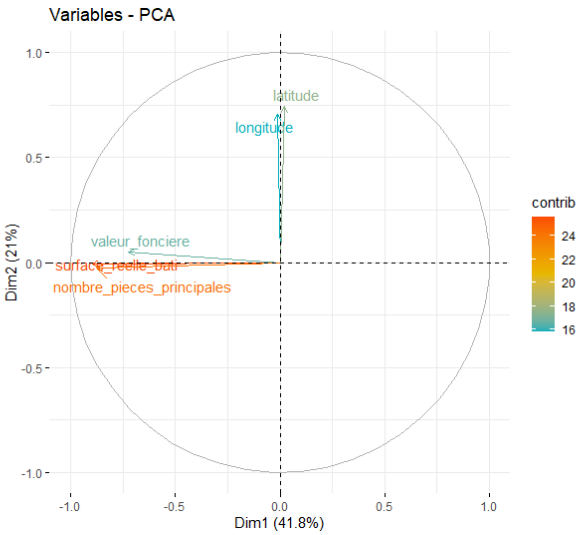
Nous avons appliqué cette fois-ci l'analyse en composantes principales sur le département de la Meuse .





Les deux premières composantes principales capturent 62.8% de l'information ainsi nous avons affiché la répartition des individus projetés dans l'espace d'arrivée colorés.
Ensuite, nous avons projeté les cinq variables dans l'espace d'arrivée.

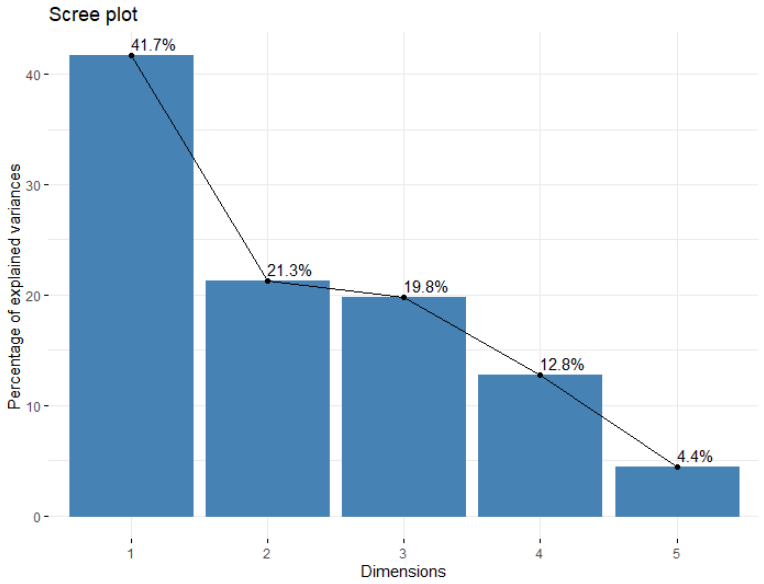
Le graphique des variables montre qu'il y a une forte corrélation entre les trois variables (la valeur foncière, la surface réelle du bâtiment et le nombre de pièces principales), ce qui s'explique par le fait qu'il y a davantage de zones rurales que de zones urbaines et qu'il y a donc une stabilité dans la corrélation entre ces trois variables.

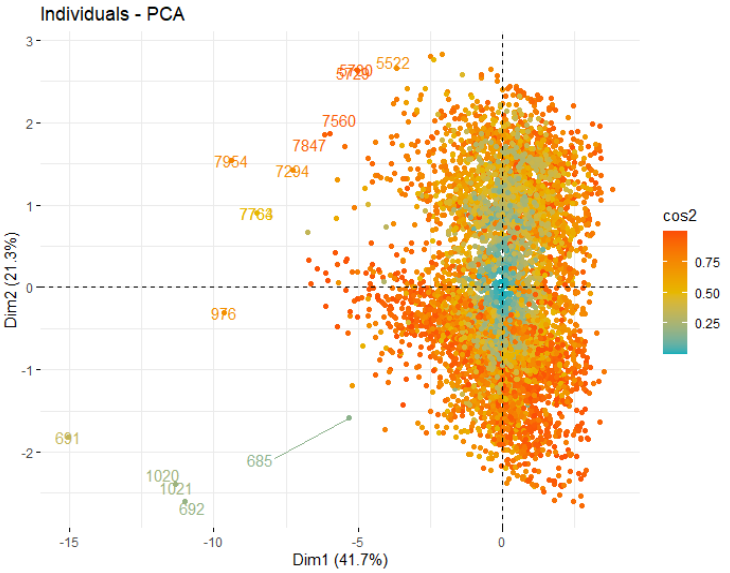


Nous avons aussi réalisé le graphique des boxplots en Meuse qui montre une grande dispersion de la valeur foncière pour une superficie comprise entre 130 et 194 et une corrélation positive entre les deux variables valeur foncière et surface réelle bâti .
les deux variables tendent à augmenter ensemble
(Voir figure 3 en Annexe) .

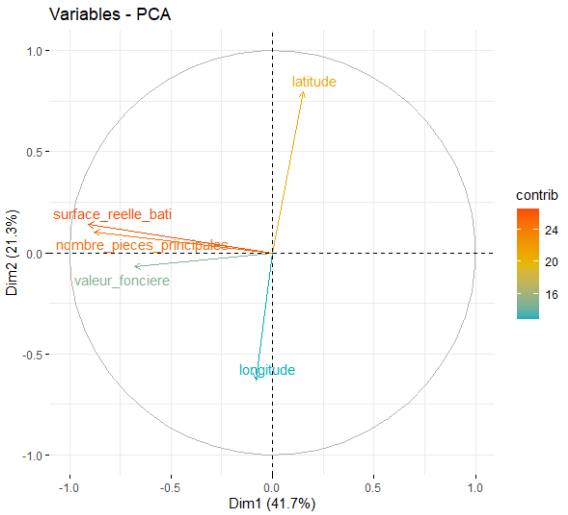
3) Département de La Loire Atlantique

Nous avons également appliqué l'analyse en composantes principales au département de La Loire Atlantique .
Le graphique ci-dessous du screen plot montre que les deux premières composantes principales capturent 63% de l'inertie de jeu de données et donc on projette les individus dans l'espace d'arrivée composé de ces deux composantes principales .





Nous avons projeté aussi les cinq variables quantitatives dans l'espace d'arrivée ,le graphique des variables ci-contre montre qu'il y a une forte corrélation entre les trois variables valeur foncière , surface réelle bâti et nombre pièces principales.



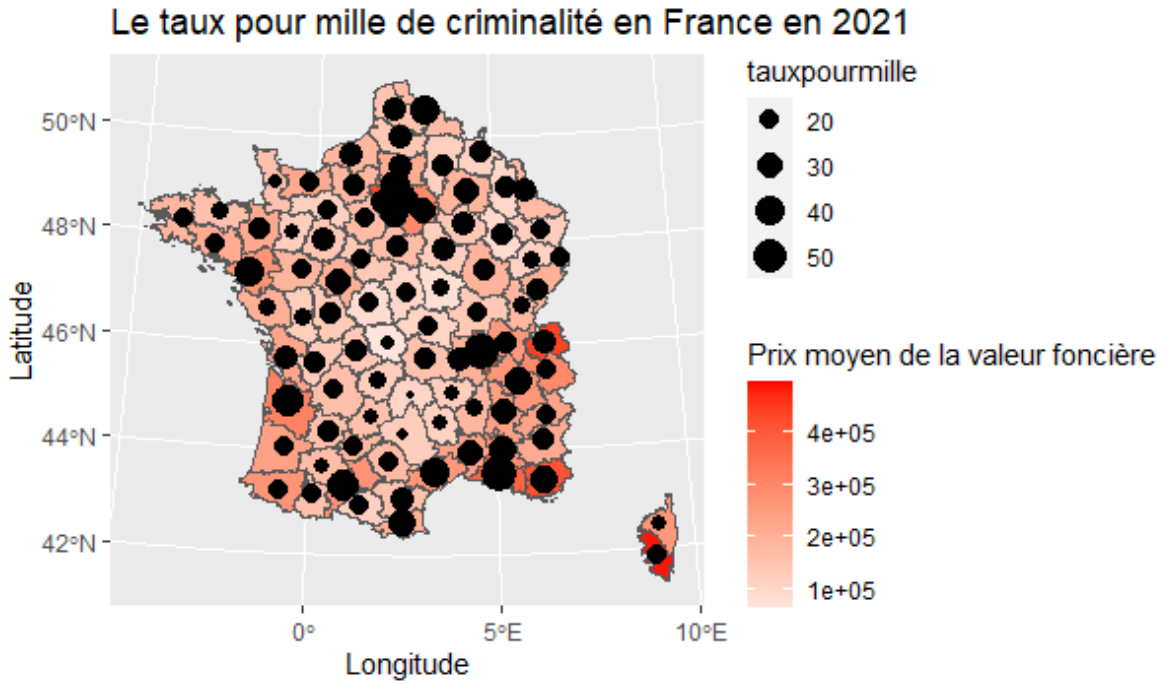
La figure 4 des boxplot **en Annexe** montre qu'il y a une grande dispersion de la valeur foncière pour une surface comprise entre 163 et 257.
Les deux variables valeur foncière et surface réelle bati tendent à augmenter ensemble .

IV) Études Comparatives

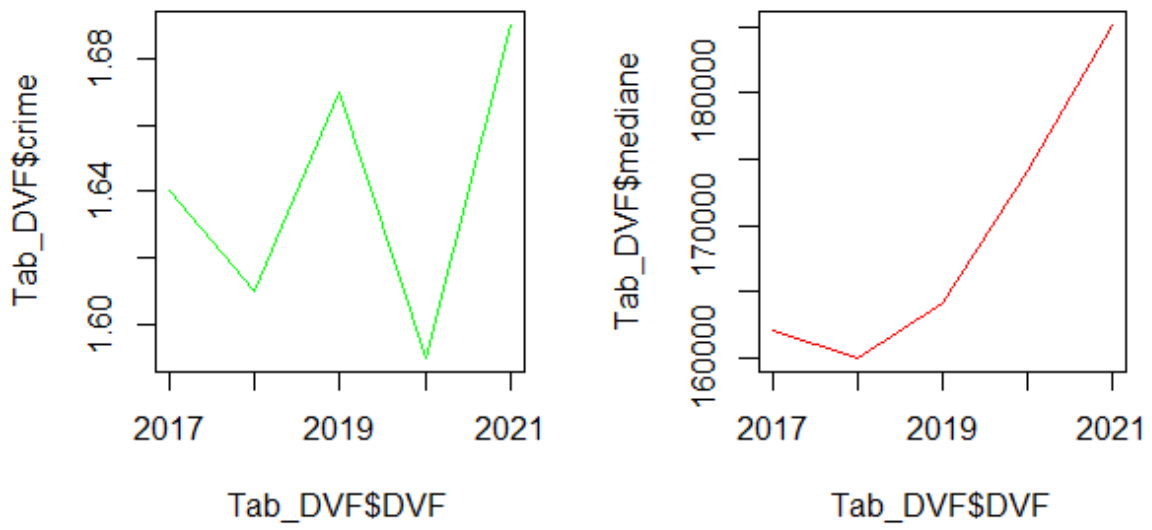
L'évolution de la valeur foncière en France est évidemment liée à des critères sociaux et économiques. En effet, nous avons essayé dans cette partie de faire des études de corrélation entre le taux de criminalité et le nombre de demandeurs d'emploi en France.

A) Lien avec la criminalité en France

Les données de criminalité sont fournies par Ministère de l'Intérieur et des Outre-Mer et sont fournies sur le site de data-gouv. Il y est renseigné les différentes catégories criminelles par années et le taux pour mille par départements. Afin d'avoir une vue globale sur la répartition du taux de criminalité en France, nous l'avons superposé avec la valeur foncière moyenne par département.



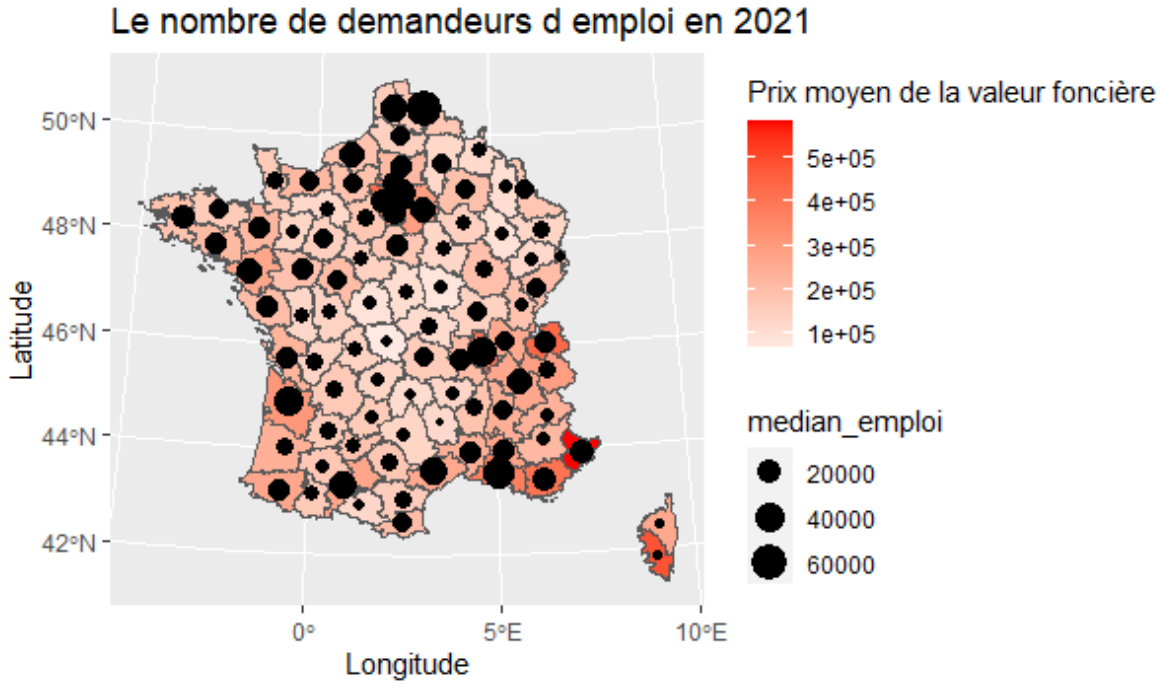
De prime abord, il est possible de confirmer que le taux de criminalité est globalement plus fort dans les départements où le prix moyen de l'immobilier est plus élevé.

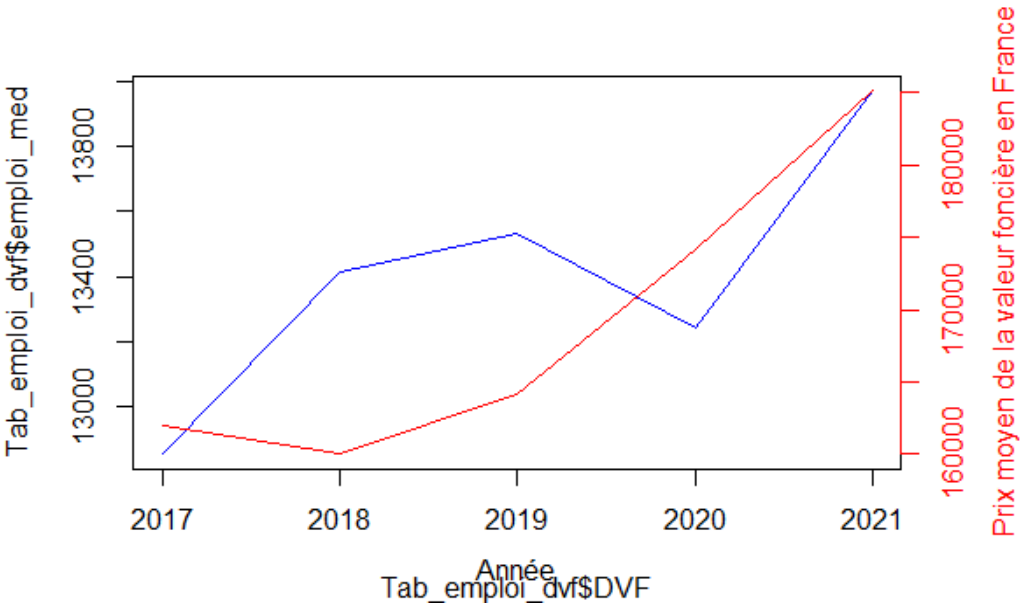


En effet, il est possible de confirmer que la valeur foncière augmente depuis 2018 (graphe de droite) mais le taux de criminalité ne dépend pas des années, malgré le pic en 2020 sûrement dû à la période du COVID (graphe de gauche). Une échelle plus locale au niveau d'un département en particulier selon les années aurait été plus intéressante à analyser.

B) Lien avec le nombre de demandeurs d'emploi en France

Un autre critère qui paraît pertinent dans l'évolution de la valeur foncière en France est le nombre de demandeurs d'emploi en France.





C'est en effet ce que l'on peut constater sur le graphique. En rouge, la courbe qui représente le fait que la valeur foncière augmente depuis 2018 et, en bleu, le nombre de demande d'emploi augmente aussi de 2017 à 2019 et de 2020 à 2021. La baisse soudaine correspond probablement à la période du COVID.

V) Apprentissage

Cette base de donnée contient de très nombreuses valeurs dont le type du local n'est pas renseigné (cf annexe ###). La base DVF contient de nombreux champs, il peut être intéressant de faire de l'apprentissage machine afin d'essayer d'étiqueter les biens par le type du local.

Ces locaux sont répartis en 4 catégories : Maison, Appartement, Local industriel ou commercial et Dépendances. A partir des 40 champs renseignés nous n'en avons gardé que ### qui paraissaient pertinents pour le programme à apprendre. Nous avons laissé de côté par exemple l'id mutation ou le nom de la parcelle car aucun lien ne relie ces données là au type de local à priori.

Maintenant certaines variables apparaissent très peu souvent mais peuvent déterminer sûrement un type de local. On les passe alors en valeurs muettes (dummy variable). Nous l'avons fait pour les surfaces carrez des lots i.

Nous avons ensuite créé le jeu de données d'apprentissage et le jeu de données test en séparant simplement les biens étiquetés et ceux non étiquetés.

Un filtre supplémentaire a été appliqué sur les données d'apprentissage, celui de supprimer les lignes qui n'ont pas de valeurs remplies pour les champs de valeur foncière, ###

Pour cet apprentissage nous avons utilisé la bibliothèque randomForest.

Voici en annexe, si je retrouve, la table de confusion de sortie après ce premier essai.

C'est pas mal mais pas non plus incroyable.

Une caractéristique importante de DVF est que pour une transaction donnée plusieurs lignes apparaissent cf annexe ###. Très peu de champs varient pour ces mêmes bien : nature_culture, nature_culture_special et surface_terrain. Une idée serait d'essentialiser plusieurs lignes en une. Nous avions d'abord pensé à sommer les surfaces terrains sauf que vu l'absence de surface_terrain dans le jeu de données test ça ne servirait à rien. On a alors simplement ajouté un champs n qui compte le nombre de lignes pour une même transaction. Une transaction unique est définie par une id_parcelle unique et une date_mutation unique. Car plusieurs bien se situent sur une même parcelle : camping ou lotissement par exemple.

Nous avons fait retourné l'algorithme avec ce champ supplémentaire mais ça n'a pas du tout aidé, les résultats donnent des taux d'erreur proche de 80 %. C'est extrêmement nul.

Nous avons donc gardé uniquement la ligne principale de chaque groupe de bien correspondant à une même transaction. Nous avons gardé la première ligne qui correspond au champs nature culture sol, type identique pour tous les biens.

```
> model

Call:
 randomForest(formula = type_local ~ ., data = new_donnees_sans_doublon,      ntree = 400)
               Type of random forest: classification
                     Number of trees: 400
No. of variables tried at each split: 3

        OOB estimate of  error rate: 12.38%
Confusion matrix:
                                    Appartement Dépendance Local industriel. commercial ou assimilé Maison class.error
Appartement                                3332         79                                      24   3070 0.487778632
Dépendance                                 1057       2879                                       9   8050 0.759983326
Local industriel. commercial ou assimilé    661        236                                    1098   9086 0.900911470
Maison                                      646        641                                     134 160444 0.008778921
```

Résultat sur l'année 2017

Globalement les résultats ne sont pas satisfaisants mais ce qui nous intéresse le plus est la classification des maisons. Ce classifieur a un excellent rappel de 99,12% et une précision de 90% pour les maisons.

Pistes d'amélioration:

-Apprendre avec le champ nature culture spécial et garder uniquement les attributs qui reviennent le plus souvent. Car limitation de la fonction randomForest sur les valeurs qualitatives: impossibilité d'en avoir plus de 53. Or le champ nature culture spéciale en compte 85.

-Étudier un autre algorithme d'apprentissage comme SVM.

-Trouver plus d'outlier car le jeu de données de base est très grand.

-Agréger avec les autres années pour avoir un modèle encore plus robuste.

-Faire la phase de test, pas eu les ressources temporelles pour le faire.

-Comparer les résultats à un classifieur naïf.

VI) Conclusion

Pour conclure, nous avons effectué une étude comparative entre 3 départements ayant des caractéristiques différentes. D'une part, nous avons pu constater que le département le moins attractif, il n'y a seulement des critères sur le type de l'immobilier qui joue sur son prix. D'autre part, les départements attractifs et dynamiques, les autres critères tels que la localisation vont influer sur le prix de l'immobilier. Pour analyser plus en détail ces différences, il serait pertinent de travailler sur une échelle plus locale.

Ainsi, les perspectives de ce projet seraient de travailler sur d'autres études comparatives, notamment d'autres critères sociaux et économiques, mais aussi de travailler sur des échelles locales sur plusieurs années.

VII) Annexe

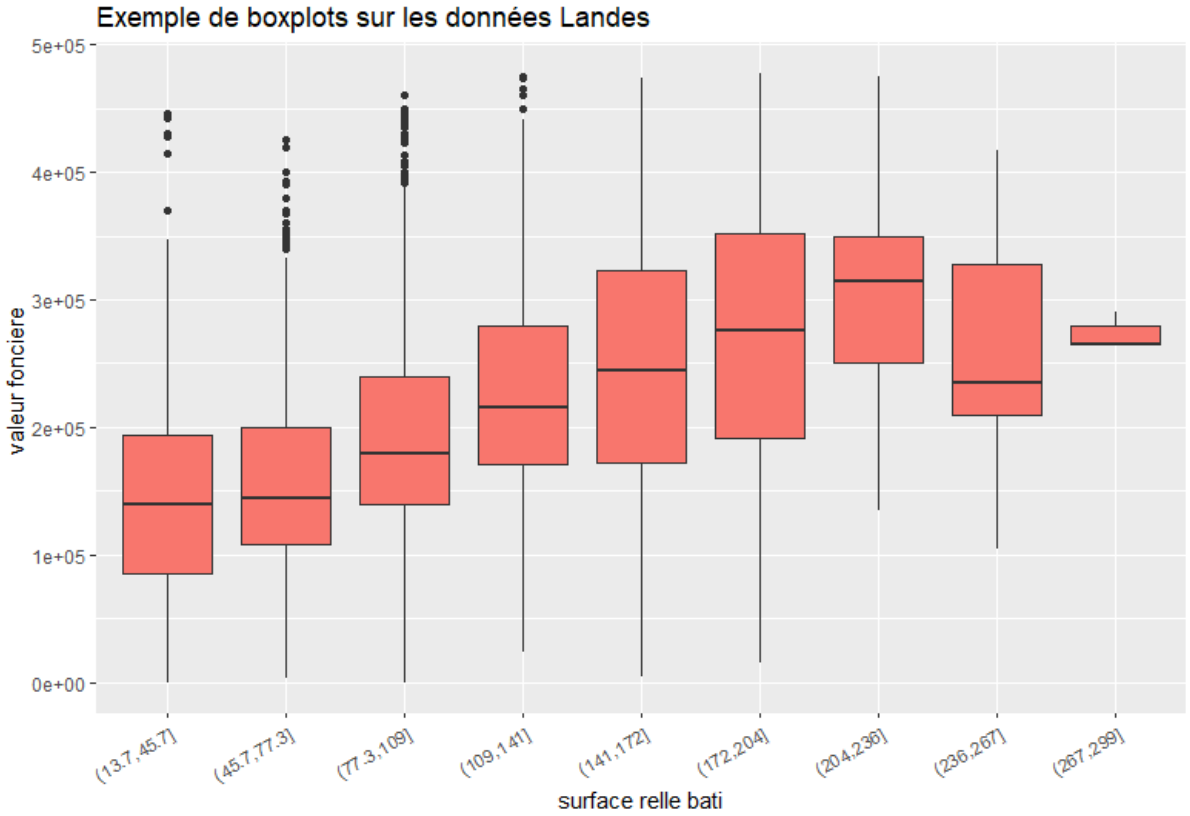
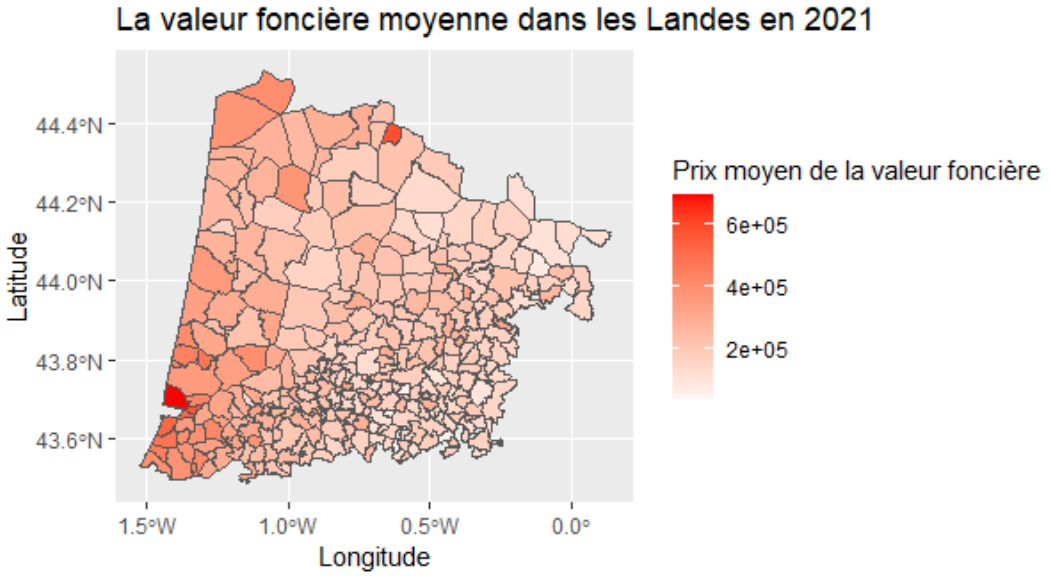


Figure 1 : Boxplots sur les données des Landes



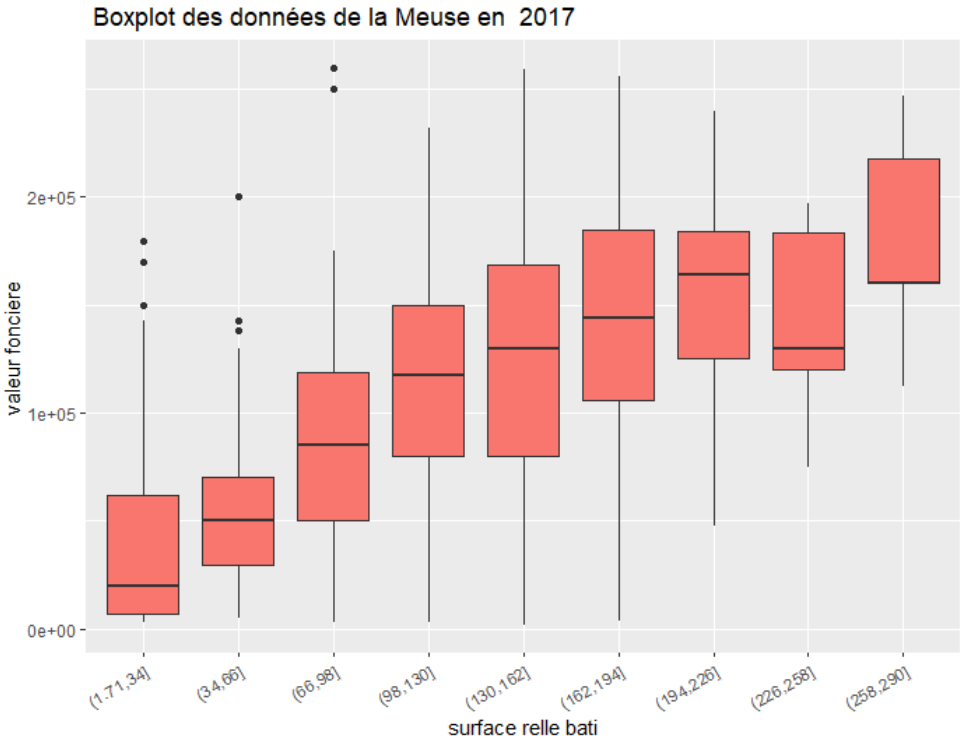


Figure 3: Boxplots sur les données de La Meuse

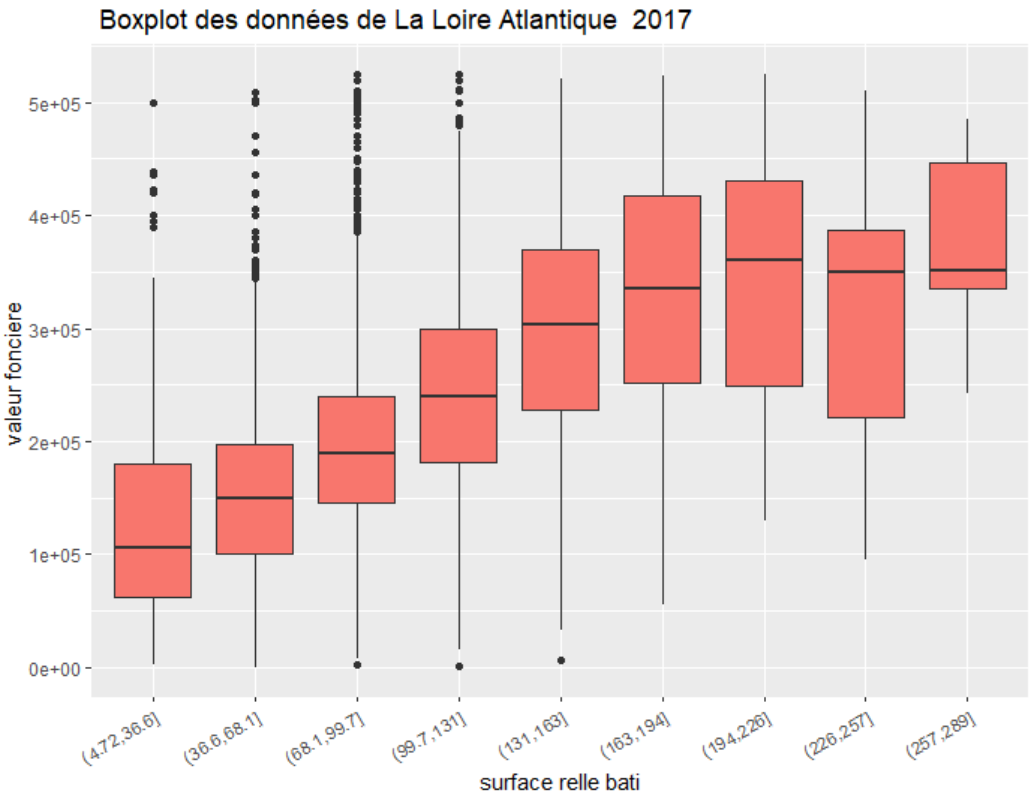


Figure 4 : Boxplots sur les données de La Loire Atlantique

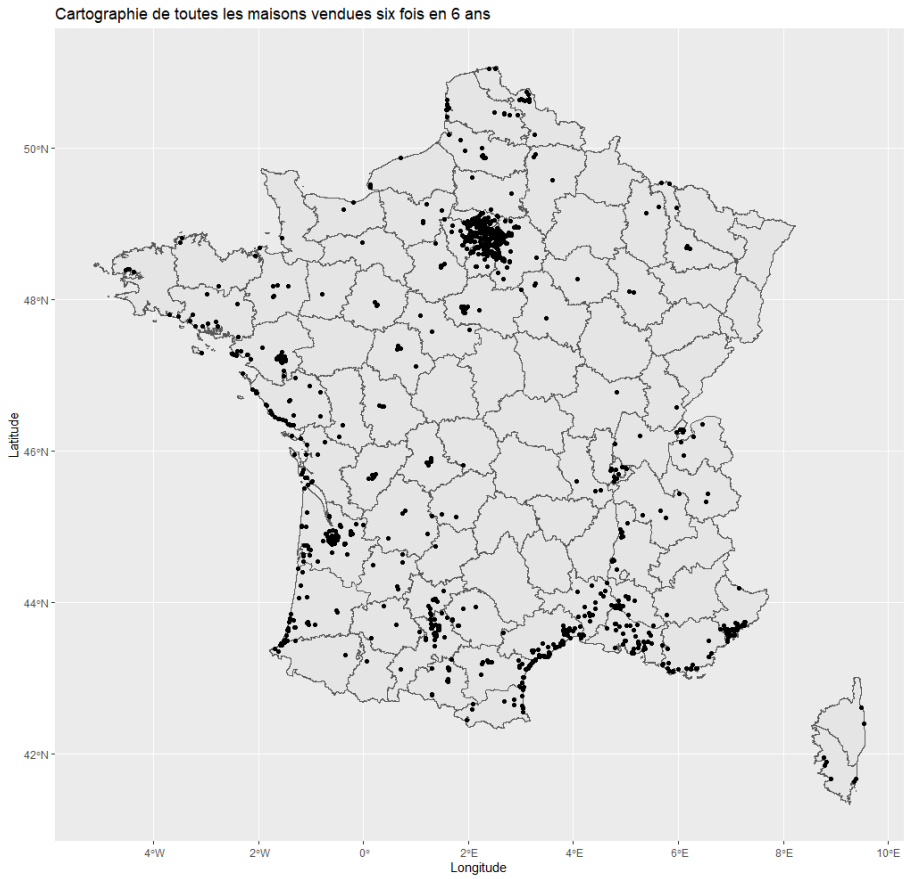


Figure 5 : Carte des maisons vendues au moins six fois ces six dernières années (France métropolitaine)

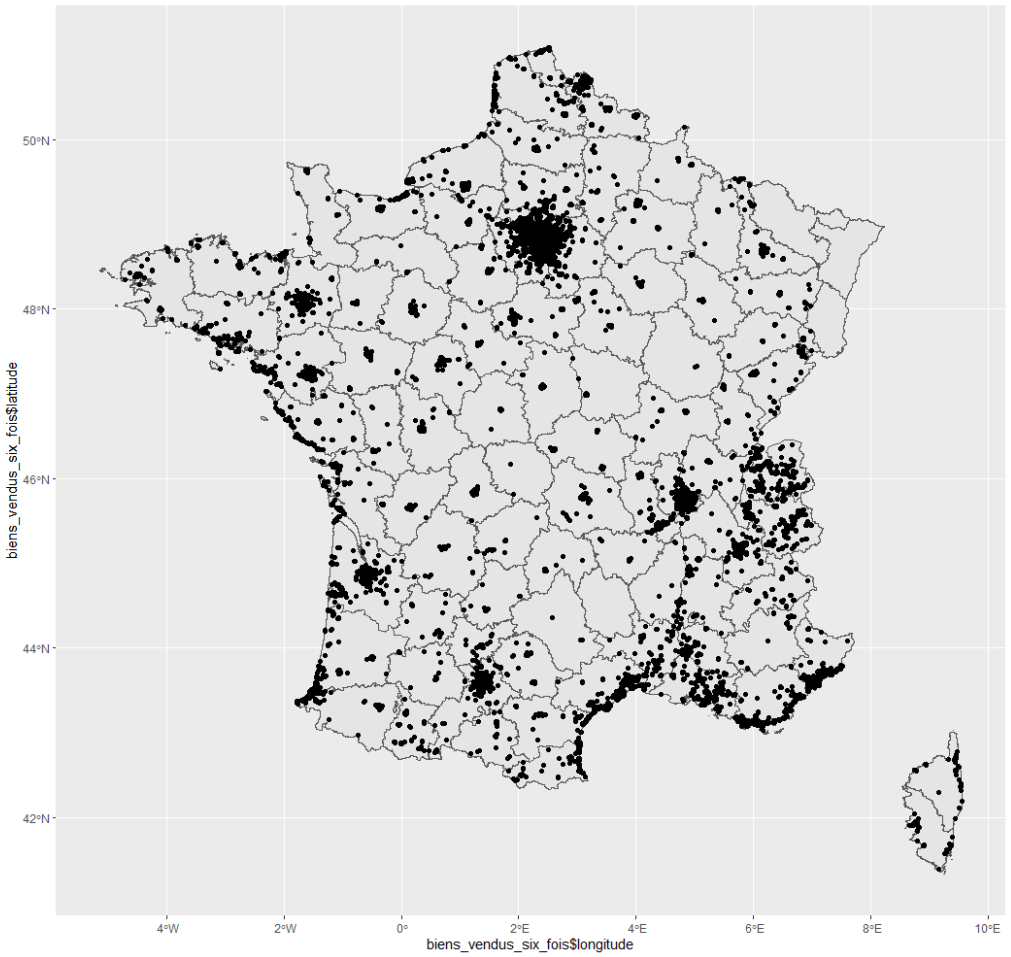


Figure 6 : Carte des biens mutés au moins six fois ces six dernières années (France métropolitaine)

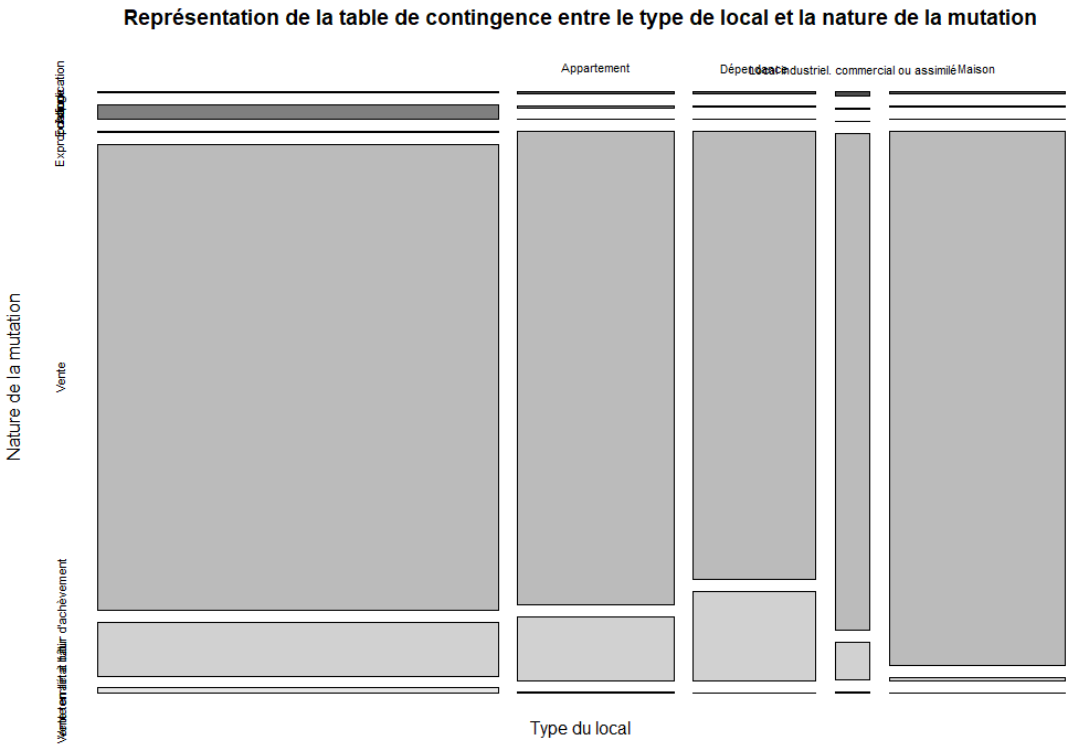


Figure 7 :Table de contingence entre le type du local et la nature de la mutation